Three Essays in Micro-Econometrics

Author: Tao Yang

Boston College

The Graduate School of Arts and Sciences

Department of Economics

THREE ESSAYS IN MICRO-ECONOMETRICS

a dissertation

by

TAO YANG

submitted in partial fulfillment of the requirements

for degree of

Doctor of Philosophy

December 2015

DISSERTATION COMMITTEE

PROF. ARTHUR LEWBEL (Chair)

PROF. STEFAN HODERLEIN

PROF. ZHIJIE XIAO

THREE ESSAYS IN MICRO-ECONOMETRICS

TAO YANG

Abstract

My dissertation is composed of three chapters. The first chapter is on the asymptotic trimming and rate adaptive inference for heavy-tail distributed estimators. The second chapter is about the identification of the Average Treatment Effect for a two threshold model. The last chapter is on the identification of the parameters of interest in a binary choice model with interactive effects.

Contents

Chapter 1

Asymptotic Trimming and Rate Adaptive Inference for Endogenous Selection Estimates

1.1 Introduction

Some common estimators in econometrics involve heavy-tailed distributions, meaning that second moments are infinite or do not exist. This is sometimes due to the structure of the estimator, and sometimes due to the presence of heavy tailed error terms (examples of both are given below). Heavy tailed estimators tend to be volatile, because of the presence of large valued observations that appear as outliers. Including or excluding a small number of these outliers may dramatically change the estimate. Making things even worse, the unbounded second moment renders inference after estimation extremely difficult, i.e., the standard central limit theorem (CLT) cannot be used.

One way to overcome the heavy-tails problem is to trim out some of those large values. However, heavy-tailed estimators are often very sensitive to the exact amount that one trims. If we trim too much, the estimator may be greatly biased due to the loss of highly informative observations, while if we trim too little, the estimator will still have high variance and possibly not be asymptotically normal. Just like the Goldilocks principle, to have the "best" estimate, we need to trim appropriately. The meaning of "best" in the current context is two fold: attaining the fastest rate of convergence possible while maintaining asymptotic normality. In this paper, we propose a general approach to deal with trimming to achieve this goal. In the application of our approach, an optimal numerical value for the trimming parameter is determined, not just an optimal rate.

Suppose we want to estimate a quantity μ with an estimator $\widehat{\mu}$ that can be represented as

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} W_i \mathbb{I}\left(-\gamma_n' \leq V_i \leq \gamma_n\right) + o_p\left(n^{-\frac{1}{2}}\right), \tag{1.1.1}$$

where observations are i.i.d., W_i is either observed or is the influence function of the estimator, and the second moment of W_i is infinite. To deal with the heavy-tailed W_i, we trim based on a variable V_i (V_i could be W_i itself) with positive trimming parameters γ_n' and γ_n, such that the estimator after trimming has finite second moment. To estimate μ consistently, we employ asymptotic trimming, i.e., the trimming parameters γ_n', γ_n go to infinity as n goes to infinity. Let x_{ni} denote $W_i \mathbb{I}\left(-\gamma_n' \leq V_i \leq \gamma_n\right)$, and let the bias and variance terms of the estimator be $\mathcal{B}_n = \mathbb{E}\left(x_{ni}\right) - \mu$ and $\sigma_n^2 = \text{var}(x_{ni})$ respectively.

Many important estimators in econometrics are in the form of equation (1.1.1). These include the following.

1. Density weighted estimators. These are estimators that are averages weighted by

the inverse of the density function of some variables. One well known example is Hardle and Stoker's (1989) average derivative estimator, and others are some special regressor estimators by Lewbel (1998, 2000, 2007). These estimators generally trim out observations where the weighted density function is close to zero.

2. Propensity score weighted average treatment effect estimation. The denominator of the average treatment effect (ATE hereafter) estimator in Hahn (1998) is the propensity score of some control variables. The second moment of this estimator is generally unbounded unless the propensity score is bounded away from zero or one. This then requires trimming out observations of the propensity score that are too close to zero or one.

3. Identification at infinity. To estimate the intercept term in a selection model, Heckman (1990) and Andrews and Schafgans (1998) propose only using those observations for which the probability of selection is close to one. They therefore trim out observations based on the probability of selection.

4. Time series models with heavy tails errors. Asymmetry and heavy tails are empirically documented in a wide range of financial, macroeconomic and actuarial time series, including exchange rate and asset price fluctuations, and in insurance claims (Mandelbrot 1963, Campbell and Hentschel 1992, Engle and Ng 1993, Embrechts et al 1997, Finkenstadt and Rootzen 2003). Trimming may therefore be needed to stabilize estimates of time series models with thick tailed data like these.

5. Microeconomic heavy-tailed data. Microeconomic data that have been shown to possess heavy tails include auction bids (Hill and Shneyerov 2013), birth weights (Cher-

nozhukov and Fernandez 2011) and network traffic (Resnick 1997). Estimators that entail averaging such data will therefore require trimming.

We show that for estimators in the form of equation (1.1.1) with heavy-tailed W_i, there exist two cases, which we refer to as the "nice" world and the "ugly" world. We show that in the "nice" world, there exists a value for the trimming parameters that gives $\widehat{\mu}$ the fastest possible convergence rate (which may be slower than root-n), and for this trimming the CLT holds and $\sqrt{\frac{\sigma_n^2}{n}}\mathcal{B}_n = O(1)$.

In contrast to the nice world, in the "ugly" world standard inference (such as t-tests or z-tests) does not work either because the CLT fails or because the bias term dominates the limiting distribution when CLT holds. Dominance of the bias term makes standard confidences intervals potential fail to cover the true value, while the failure of CLT makes inference extremely difficult, e.g., in many case even the existence of an asymptotic distribution may unknown.

It is therefore important to know which world we are in for any given application. We give a general method to tell if the world is nice or ugly, and in the nice case, show how to choose the trimming parameters to have the fastest convergence rate (which may be slower than root-n) while still having the CLT hold.

Our procedure consists first of applying the Lindeberg-Feller central limit theorem (see Theorem 1.2.1 in Section 1.2.1) for the asymptotic normality of arrays $\{x_{ni}\}_{i=1}^n$. Under some weak regularity conditions, this CLT says that asymptotic normality holds if and only if the Lindeberg condition (equation 1.2.1) is satisfied. We first look for the largest possible set of values of the trimming parameters γ_n' and γ_n for which the Lindeberg condition holds.

If asymptotic normality holds, then

$$\sqrt{n}\left(\frac{\widehat{\mu}-\mu-\mathcal{B}_n}{\sigma_n}\right)\xrightarrow{d}N\left(0,1\right).$$

Our procedure next finds the values of the trimming parameters, from previously obtained set of values for which the Lindeberg condition holds, that minimize the rate of Root Mean Squared Errors (RMSE) subject to $\sqrt{\frac{\sigma_n^2}{n}}\mathcal{B}_n = O\left(1\right)$ and thereby achieve this fastest rate of convergence.

If this minimizing value of the trimming parameters exists, then we are in the nice world, and these are optimal values of γ_n' and γ_n for estimation and inference. Otherwise we are in the ugly world and standard inference is not possible.

This procedure demonstrates the importance of finding the largest possible set of values for the trimming parameters that can satisfy the Lindeberg condition. The difficulty in doing so stems from the fact that the expression of the Lindeberg condition is complicated.

Papers including Bickel (1982), Manski (1984), Robinson (1988), and Hardle and Stoker (1989) use asymptotic trimming to handle boundary bias in nonparametric estimation, which is different from the goal here. Another strand of literature, including Hill (1975) and Csorgo, Haeusler and Mason (1988 a, b) apply asymptotic trimming to averages of series that are in the non-normal domain of attraction. In the above notation, this literature assumes that

$$P\left(|W_i|>\gamma\right)=c_1\gamma^{-c_2}\left(1+o\left(1\right)\right), \tag{1.1.2}$$

for some $c_1>0$, $c_2\in(1,2]$. Condition (1.1.2) is the definition of a stable distribution, implying that for any $c>c_2$, $\mathbb{E}\left(|W_i|^c\right)=\infty$ and the convergence rate of $\frac{1}{n}\sum_{i=1}^n W_i-\mathbb{E}\left(W\right)$

is n^{1-1/c_2}. For more about stable distributions, see Samorodnitsky and Taqqu (1997). Using this approach, Chaudhuri and Hill (2013) do asymptotic trimming for propensity score weighted ATE estimation and Hill and Renault (2010) do asymptotic trimming for time series models with heavy-tailed errors. However, the assumption of a stable distribution is rather restrictive. Moreover, the convergence rate of the estimation of c_2 as needed for selecting trimming parameters is extremely slow, i.e., $\log(n)$. In contrast, our approach in this paper makes no comparable modeling assumption about the tails of W_i.

Some papers address the heavy tails problem by trimming a fixed portion of extreme observations. For example, in the context of the average treatment effect model, see Potter (1993), Frolich (2004), Lee, Lessler, and Stuart (2011) and Chaudhuri and Min (2012). However, fixed trimming like this leads to inconsistency in most cases.

Two related papers to ours are Andrews and Schafgans (1998) and Khan and Tamer (2010). The former deals with asymptotic trimming for the intercept term in a selection model, while the latter focuses on trimming of the weighted ATE estimator (Hahn 1998) and of special regressor binary choice model estimation (Lewbel 2000). Both papers also use the Lindeberg Feller CLT as the main tool, but they either use sufficient conditions for the Lindeberg condition to choose γ_n', γ_n, or simply assume the Lindeberg condition holds, or assume specific distributions on unobserved error terms. In contrast, in this paper we deal with the Lindeberg condition directly. Moreover, we relax assumptions on the distributions of unobserved error terms.

Another related paper is Khan and Nekipelov (2014), which gets the uniform inference procedure around the boundary of the regular and irregular identification (whether endogeneity exists or not) of the endogenous selection model, using the stable distribution

approach.

To demonstrate how our approach works, we derive a characterization of the required Lindeberg condition for a class of estimators that are weighted with inverse density functions. We then apply our method to the special regressor estimator in an endogenous selection model, which is an example of an inverse density weighted estimator.

For illustration purposes, consider the following simple endogenous selection model (our later application will be a richer model that includes covariates):

$$Y = Y^* D, \tag{1.1.3}$$

$$D = \mathbb{I}\left(V - U \geq 0\right), \tag{1.1.4}$$

where $\mathbb{I}\left(\cdot\right)$ is the indicator function equalling one when the argument inside is true and zero otherwise, Y is an observed outcome, D is an observed treatment indicator, and U is an unobserved confounder which is possibly correlated with the unobserved latent outcome Y^*. The goal is estimation of $\mathbb{E}\left(Y^*\right)$. In general, identification requires some variable that affects treatment but not outcomes, which in this example is an observed exogenous continuous variable V.

An example of the above model could be a wage equation. Let Y^* and Y be the true underlying and observed wage respectively. Some unobserved drive or ability measure U affects both the decision to work (D) and potential wages (Y^*). Because of this endogeneity of U, the observed average wage $\mathbb{E}\left(Y\right)$ in general differs from $\mathbb{E}\left(Y^*\right)$. The instrument V here could be $-\log\left(\text{non-labor income}\right)$, which is assumed to only affects one's desire to work but not one's wage.

Suppose we observe $\{Y_i, D_i, v_i\}_{i=1}^n$. Then a consistent estimator for $\mathbb{E}\left(Y^*\right)$ based on

Lewbel (2007) is:

$$\widehat{\mu}_n = \frac{\frac{1}{n}\sum_{i=1}^{n} \frac{D_i Y_i}{f_v(v_i)} \mathbb{I}\left(-\gamma_0 \leq v_i \leq \gamma_n\right)}{\frac{1}{n}\sum_{i=1}^{n} \frac{D_i}{f_v(v_i)} \mathbb{I}\left(-\gamma_0 \leq v_i \leq \gamma_n\right)}, \tag{1.1.5}$$

where f_v is the density function for V, γ_0 is a fixed positive number, and γ_n goes to infinity as n tends to infinity. $\frac{DY}{f_v}$ and $\frac{D}{f_v}$ corresponds to the W in equation (1.1.1). As we show later, under weak conditions, the second moments of $\frac{DY}{f_v}$ and $\frac{D}{f_v}$ do not exist. The trimming indicator gives this estimator a bounded second moment, but at the same time makes it biased. Consistency then requires that $\gamma_n \to \infty$ as $n \to \infty$.

Trimming and inference for this problem might be possible using stable distributions instead of our methodology, however, the stable distribution condition (1.1.2) is restrictive and generally unreasonable here.[1]

To make inference as easy as possible, we prove that the classic bootstrap works in the nice world, even when the convergence rate here is not the usual root-n. We generalize the identification, inference, and the trimming procedure in the above example to a semiparametric case, where there are additional covariates X and associated parameter vector β, and to the fully nonparametric case that includes covariates X and no parametric structure is imposed. We obtain a condition for f_v that indicates whether we are in the nice or ugly world. While we focus on estimation and inference for the nice world, we also consider the possibility of employing a jackknife procedure to deal with the ugly world case. We conduct a Monte Carlo analysis to check the small sample properties of our trimming procedure,

[1] One simple case where $\frac{DY}{f_v}$ in equation (1.1.5) is not distributed as a stable distribution is the following. Suppose Y^* is a constant equalling 1, $U = 0$, and V is standard normal. Then

$$\lim_{\gamma\to\infty} p\left(\frac{1\mathbb{I}\left(V>0\right)}{f_v\left(V\right)} > \gamma\right) \Big/ \gamma^{-c} = \lim_{\gamma\to\infty} \frac{\int_{\sqrt{2\log\gamma}}^{+\infty} \exp\left(-\frac{v^2}{2}\right) dv}{\gamma^{-c}} = \lim_{\gamma\to\infty} \frac{\gamma^{c-1}}{c\sqrt{2\log\gamma}} = \begin{cases} \infty, \text{ if } c > 1 \\ 0, \text{ if } c \leq 1 \end{cases},$$

where the second equality holds by L'Hopital's rule. It is not hard to see that this simple example does not satisfy the stable distribution condition.

and we apply our method empirically in a model of the gender wage gap using Malaysian data. Finally, we outline how our method might be applied to deal with other important classes of estimators.

The structure of this paper is as follows. In Section 2, we show how to choose trimming parameters and apply the method to the estimation of an endogenous selection model. In Section 3, we give a linear representation of our estimator with asymptotic trimming when f_v is nonparametrically estimated. In Section 4, we generalize our method to the semiparametric and nonparametric case when we have additional covariates. In Section 5, we check the small sample behavior of our estimator by Monte Carlo simulations. In Section 6, we apply our estimator to investigate the gender wage gap. We conclude in Section 7. In the Appendix, we discuss possible ways to deal with the ugly world case, and consider potential extensions. All proofs are in the Appendix.

We use the following notation conventions throughout this paper: upper case letters denote random variables, lower case letters denote realization; c is some constant that may vary line by line; $\equiv$ denotes definition; and the binary operator $\asymp$ denotes the same order, i.e., $a_n \asymp b_n$ means $0 < \liminf_{n\to\infty} \frac{a_n}{b_n} \leq \limsup_{n\to\infty} \frac{a_n}{b_n} < \infty$.

1.2 Trimming and Inference

1.2.1 Rate Adaptive Inference

In this subsection, we discuss the way to choose trimming parameters γ_n' and γ_n for the estimator $\widehat{\mu}$. To attain asymptotic normality, we need to choose trimming parameters such that the estimator satisfies the following Lindeberg-Feller CLT. To present the theorem formally, we let $\sigma_{ni}^2 \equiv \text{var}(x_{ni})$, though σ_{ni}^2 does not vary across i under the current i.i.d.

assumption. We let $\tau_n^2 \equiv \frac{1}{n}\sum_{i=1}^{n}\sigma_{ni}^2$.

Theorem 1.2.1 (Lindeberg-Feller CLT) *Suppose* $\{x_{ni}\}_{i=1}^{n}$ *are independent and* $\frac{\max_{i=1,\ldots,n}\{\sigma_{ni}^2\}}{n\tau_n^2} \to 0$, *then* $\frac{\sqrt{n}(\widehat{\mu}-\mu-\mathcal{B}_n)}{\tau_n} \xrightarrow{d} N(0,1)$, *if and only if, for any* $\varepsilon > 0$,

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(\frac{(x_{ni}-\mathbb{E}x_{ni})^2}{\sigma_{ni}^2}\mathbb{I}\left[\frac{(x_{ni}-\mathbb{E}x_{ni})^2}{\sigma_{ni}^2} > n\varepsilon\right]\right) = 0. \tag{1.2.1}$$

Condition $\frac{\max_{i=1,\ldots,n}\{\sigma_{ni}^2\}}{n\tau_n^2} \to 0$ in above theorem means that no single observation contributes a significant portion in total variance of the estimator. This holds in most econometrics models, such as those with i.i.d. data. Under independence and this weak condition, the Lindeberg-Feller CLT states that asymptotic normality holds, if and only if equation (1.2.1) is satisfied. Equation (1.2.1) is the Lindeberg condition. By this theorem, for the estimator $\widehat{\mu}$, we only need to check the Lindeberg condition to see if asymptotic normality holds or not.

Under the i.i.d. assumption, the Lindeberg condition can be further simplified to

$$\lim_{n\to\infty}\mathbb{E}\left(\frac{(x_{ni}-\mathbb{E}x_{ni})^2}{\sigma_{ni}^2}\mathbb{I}\left[\frac{(x_{ni}-\mathbb{E}x_{ni})^2}{\sigma_{ni}^2} > n\varepsilon\right]\right) = 0.$$

Define a set for γ_n' and γ_n

$$\Psi \equiv \bigcap_{\varepsilon>0}\left\{(\gamma_n',\gamma_n)\,\middle|\,\lim_{n\to\infty}\mathbb{E}\left(\frac{(x_{ni}-\mathbb{E}x_{ni})^2}{\sigma_{ni}^2}\mathbb{I}\left[\frac{(x_{ni}-\mathbb{E}x_{ni})^2}{\sigma_{ni}^2} > n\varepsilon\right]\right) = 0\right\}. \tag{1.2.2}$$

By the Lindeberg-Feller CLT, asymptotic normality holds if and only if we choose trimming parameters from Ψ.

Define a set

$$\Gamma = \left\{ \gamma_n', \gamma_n \,\middle|\, (\gamma_n', \gamma_n) \in \Psi \text{ and } \sqrt{\frac{n}{\sigma_n^2}} B_n = O(1) \right\}.$$

If we choose trimming parameters from set Γ, then we have asymptotic normality and inference is possible. If Γ is not empty, then we say we are in the nice world. Otherwise we are in the ugly world: either the CLT fails or the bias term is the dominant term when CLT holds. In the ugly world, standard t-tests or z-tests cannot give valid inference when the CLT holds because of the dominance of the bias term and standard t-tests or z-tests are not available when the CLT doesn't hold. Moreover, when normality fails, alternative inference procedures like the bootstrap are often invalid, e.g., see Khan and Nekipelov (2014). Consequently, inference is difficult in the ugly world.

Once we know we are in the nice world, the next step is to choose (γ_n', γ_n) from Γ to minimize the rate of RMSE $\sqrt{\mathcal{B}_n^2 + \sigma_n^2/n}$. In this way, we have asymptotic normality and the fastest convergence rate while inference is possible. The following is a formal definition of the nice and ugly world.

Definition 1.2.2 *Suppose* $\{x_{ni}\}_{i=1}^n$ *are independent and* $\frac{\max_{i=1,\ldots,n}\{\sigma_{ni}^2\}}{n\tau_n^2} \to 0$. *We say we are in the nice world, if* Γ *is not empty. Thus, in the "nice" world, we could obtain the following from some trimming parameters in the set* Ψ*:*

$$\sqrt{\frac{n}{\tau_n^2}} (\widehat{\mu} - \mu - \mathcal{B}_n) \xrightarrow{d} N(0,1),$$

and $\sqrt{\frac{n}{\tau_n^2}} \mathcal{B}_n = O(1)$. *Otherwise we say we are in the "ugly" world; for any* $(\gamma_n', \gamma_n) \in \Psi$, *we have* $\limsup \sqrt{\frac{n}{\tau_n^2}} \mathcal{B}_n = \infty$.

The key to achieve our goal in this paper is to know Ψ. However, the Lindeberg condition as shown in equation (1.2.1) is complicated and cannot directly be used in practice in most cases. As a result, to find the boundary between the nice and ugly world, and thereby maximize the set of models that can achieve standard inference and optimal rates, we need to find a practical way to characterize the set Ψ. In the next subsection, we find some simple conditions that are equivalent to the Lindeberg condition for a class of estimators that are weighted with inverse density functions. We then apply this result to the example of a special regressor estimator for an endogenous selection model, which uses this weighting.

1.2.2 Lindeberg Condition for Inverse Density Weighted Estimators

Here we study the Lindeberg condition for estimators that are weighted with inverse density functions. We represent those estimators as

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \frac{\varsigma_i}{f_v(v_i)} \mathbb{I}\left(-\gamma_n' \leq v_i \leq \gamma_n\right) + o_p\left(n^{-\frac{1}{2}}\right), \tag{1.2.3}$$

where f_v is the density function for V and ς_i denotes the rest of the estimator to be weighted. In the notation of equation (1.1.1), we have $W_i = \frac{\varsigma_i}{f_v(v_i)}$ and $x_{ni} = \frac{\varsigma_i}{f_v(v_i)} \mathbb{I}\left(-\gamma_n' \leq v_i \leq \gamma_n\right)$. We assume that v_i is a scalar. By some calculation, $\mathbb{E}\left(x_{ni}^2\right) = \int_{-\gamma_n'}^{\gamma_n} \frac{\mathbb{E}\left(\varsigma^2 \middle| V=v\right)}{f_v(v)} dv$. We impose the following technical assumptions.

Assumption 1 *Observations are i.i.d..*

Assumption 2 *V is continuous with support $\mathbb{R}$. For any $\gamma > 0$, $\inf\{f_v(v) \,|v \in [-\gamma, \gamma]\} > 0$, and $\lim_{v \to \pm\infty} f_v(v) = 0$.*

Assumption 3 ***(Restriction on f_v)*** *$f_v(v)$ is continuous. There exists an $\widetilde{f}_v(v)$, such*

that $\widetilde{f}_v(v) \geq f_v(v)$, $\widetilde{f}_v(v) \asymp f_v(v)$, $\widetilde{f}_v(v)$ *is monotone decreasing after some large* v *at both tails, and* $\int_{\mathbb{R}} \widetilde{f}_v(v)\,dv \leq c_b$, *where* c_b *is some positive constant.*

Assumption 4 $\mathbb{E}(x_{ni})$ *is uniformly bounded.* $\mathbb{E}\left(\varsigma^2 \middle| V = v\right)$ *is uniformly bounded.* $\mathbb{E}\left(\varsigma^2 \middle| V = v\right)$ *is either bounded away from 0, or decrease in order to zero like* f_v *in Assumption 3.*

Assumption 5 W_i *has unbounded second moment if we do not trim at both sides.*

Assumption 6 *Let* $\omega(v) \equiv \frac{\mathbb{E}\left(\varsigma^2 \middle| V=v\right)}{f_v(v)}$, *for some particular* $c_f, c_f' \in (0,1)$ *defined in equation (1.2.5),* $\limsup_{\gamma\to\infty} \frac{\omega\left((1-c_f)v\right)}{\omega(v)} < \frac{1}{1-c_f}$, *and* $\limsup_{\gamma\to-\infty} \frac{\omega\left((1-c_f')v\right)}{\omega(v)} < \frac{1}{1-c_f'}$.

Assumption 1 could be relaxed to allow heteroskedasticity as long as the conditions in the Lindeberg-Feller CLT are satisfied. Assumption 2 generally leads to irregular convergence, which motivates the need for trimming. If $f_v(v)$ is monotonically decreasing in the right tail after some large value, Assumption 3 is automatically satisfied by setting $\widetilde{f}_v(v) = f_v(v)$. $\widetilde{f}_v(v) \asymp f_v(v)$ means that $\widetilde{f}_v(v)$ can represent $f_v(v)$ in terms of decreasing rate. Assumption 3 rules out some badly behaved density functions, e.g., $\limsup_{v\to\infty} f_v(v) \geq c$, for some constant c. Assumptions 4 is mild and is made for theoretical convenience. Assumption 5 defines the heavy-tails problem. Assumption 6 says $\omega(v)$ cannot decrease too fast, and is a mild restriction in our context. Taking the right hand side as an example, by Assumption 5, we have $\int_0^{\gamma_n} \omega(v)\,dv \to \infty$. For functions $\frac{1}{v^c}, c > 0$, only those with $c \leq 1$ will let $\int_0^{\gamma_n} \omega(v)\,dv \to \infty$. $\omega(v) = \frac{1}{v}$ which is excluded by Assumption 6, but other functions with slightly thicker tails, e.g., $\omega(v) = \frac{1}{v^{1-\varepsilon}}$ for any small $\varepsilon > 0$, satisfy Assumption 6. The intuition is that the condition $\int_0^{\gamma_n} \omega(v)\,dv \to \infty$ excludes $\omega(v)$ decreasing too fast.

The following two theorems give the main results in this section.

Theorem 1.2.3 (Sufficiency) *Suppose Assumption 1 ∼ 5 hold. If*

$$\begin{aligned}
nf_v^2(\gamma_n)\int_0^{\gamma_n}\frac{\mathbb{E}\left(\varsigma^2\middle|V=v\right)}{f_v(v)}dv &\to \infty, \text{ if } \lim_{n\to\infty}\frac{\int_{-\gamma_n'}^{0}\frac{\mathbb{E}\left(\varsigma^2\middle|V=v\right)}{f_v(v)}dv}{\int_0^{\gamma_n}\frac{\mathbb{E}(\varsigma^2|V=v)}{f_v(v)}dv}=0,\\
nf_v^2(-\gamma_n')\int_{-\gamma_n'}^{0}\frac{\mathbb{E}\left(\varsigma^2\middle|V=v\right)}{f_v(v)}dv &\to \infty, \text{ if } \lim_{n\to\infty}\frac{\int_{0}^{\gamma_n}\frac{\mathbb{E}\left(\varsigma^2\middle|V=v\right)}{f_v(v)}dv}{\int_{-\gamma_n'}^{0}\frac{\mathbb{E}(\varsigma^2|V=v)}{f_v(v)}dv}=0, \qquad (1.2.4)\\
n\min\left\{f_v^2(\gamma_n),f_v^2(-\gamma_n')\right\}\int_{-\gamma_n'}^{\gamma_n}\frac{\mathbb{E}\left(\varsigma^2\middle|V=v\right)}{f_v(v)}dv &\to \infty, \text{ otherwise.}
\end{aligned}$$

then the Lindeberg condition (1.2.1) holds.

Theorem 1.2.4 (Necessity) *Let Assumption 1 ∼ 6 hold. Suppose for density function f_v, there exist constants $0<c_f,c_f'<1$, $a_f,a_f'>1$, such that*

$$\begin{aligned}
\frac{f_v^2\left(\left(1-c_f\right)a_f\gamma\right)\int_0^{a_f\gamma}\frac{1}{f_v(v)}dv}{f_v^2(\gamma)\int_0^{\gamma}\frac{1}{f_v(v)}dv} &= O(1), \text{ as } \gamma\to\infty, \qquad (1.2.5)\\
\frac{f_v^2\left(\left(1-c_f'\right)a_f'\gamma\right)\int_{a_f'\gamma}^{0}\frac{1}{f_v(v)}dv}{f_v^2(\gamma)\int_{\gamma}^{0}\frac{1}{f_v(v)}dv} &= O(1), \text{ as } \gamma\to-\infty.
\end{aligned}$$

Then if the Lindeberg condition (1.2.1) holds, either condition (1.2.4) holds or the condition (1.2.4) can give the fastest rate of γ_n.[2]

Theorem 1.2.3 gives a sufficient condition for the Lindeberg condition. The regularity condition (1.2.5) is needed to let the sufficient condition also be the necessary condition or to give the fastest rate of trimming parameters. Lemma 1.2.5 shows that virtually all standard distributions (e.g. Cauchy, student t, exponential, normal) satisfy this condition. Only very thin tailed distributions like the extreme value distribution fail this regularity condition. It

[2]The meaning of the fastest rate of γ_n is as follows. Suppose $(-\gamma_n',\gamma_n)$ from condition (1.2.4) can have the Lindeberg condition hold. $\left(-a_{n,i}'\gamma_{n,i}',a_{n,i}\gamma_{n,i}\right)$ is a sub-series where $\left\{a_{n,i}'\right\}$ or $\{a_{n,i}\}$ goes to infinity, and $\left(-a_{n,i}'\gamma_{n,i}',a_{n,i}\gamma_{n,i}\right)$ fails condition (1.2.4). Then $\left(-a_{n,i}'\gamma_{n,i}',a_{n,i}\gamma_{n,i}\right)$ fails the Lindeberg condition.

is therefore usually reasonable to impose this f_v not too-thin-tailed restriction. Note that condition (1.2.4) is sufficient and close to necessary for the Lindeberg condition, and the iff condition set of trimming parameters to the Lindeberg condition is a slight expansion from condition (1.2.4). As shown in Section 1.2.4 by Theorem 1.2.9, when we have a little more structure on ς_i we can strengthen the current near iff condition to an actual iff condition, and can do so under an even more general regularity condition that allows for extremely thin-tailed distributions.

Lemma 1.2.5 *If $f_v(v)$ decays in the right tail at the same order as $\frac{1}{v^{1+c}}$, $v^{c_1}\exp(-v^{c_2})$, $\exp(-v^c)$, v^{-v^c}, for any $c, c_1, c_2 > 0$, condition (1.2.5) is satisfied. If $f_v(v)$ decays in the right tail at the same order as $\exp(-\exp(v^c))$, for any $c > 0$, then condition (1.2.5) fails.*

Some applications only need one sided asymptotic trimming. Suppose we do fixed trimming or no trimming at left hand side, so the trimming indicator becomes $\mathbb{I}(-\gamma_0 \leq v_i \leq \gamma_n)$, where γ_0 is a fixed positive number or infinity. Following the same line analysis, we get the following corollary.

Corollary 1.2.6 *Under the same assumptions as in Theorem 1.2.3 and 1.2.4, for the estimator (1.2.3) with the one-side trimming indicator $\mathbb{I}(-\gamma_0 \leq v_i \leq \gamma_n)$, the sufficient and near necessary condition (in the sense of Theorem 1.2.4) to the Lindeberg condition becomes*

$$nf_v^2(\gamma_n)\int_{-\gamma_0}^{\gamma_n}\frac{\mathbb{E}\left(\varsigma^2\middle|V=v\right)}{f_v(v)}dv \to \infty. \tag{1.2.6}$$

We next apply these results to the special regressor estimator for an endogenous selection model.

1.2.3 Identification and Asymptotic Trimming in Selection Models

The seminal papers Heckman (1976, 1979) propose two-step estimators to correct for sample selection bias. Thereafter, much work has been done on this issue, e.g., Powell (1984), Heckman (1990), Vella (1992, 1998), Ahn and Powell (1993), Wooldridge (1995), Lee (1994), Chen (1997), Honore, Kyriazidou and Udry (1997), Li and Wooldridge (2002), Abadie (2003), Das, Newey, and Vella (2003), Lewbel (2007) and many others. We apply our approach to the estimator in Lewbel (2007).

The following are our identification assumptions. We only need one-sided asymptotic trimming for the estimator here, so we weaken Assumption 2 to Assumption 10.

Assumption 7 *Observations are i.i.d. across i.*

Assumption 8 $cov\left(Y^*, U\right)$ *is finite.*

Assumption 9 $V \perp U, \mathbb{E}\left(Y^* | U, V\right) = \mathbb{E}\left(Y^* | U\right), 0 < \underline{c} \leq var\left(Y^{*2} | U, V\right) \leq \mathbb{E}\left(Y^{*2} | U, V\right) \leq \overline{c} < \infty$ *for any* U, V.

Assumption 10 V *is continuous with support* $\mathbb{R}$. *There exists a large* $\gamma_0 > 0$, *for any* $\gamma > 0$, $\inf\left\{f_v\left(v\right) | v \in [-\gamma_0, \gamma]\right\} > 0$, *and* $\lim_{v \to +\infty} f_v\left(v\right) = 0$.

Assumptions 8∼10 are the identification assumptions from Lewbel (2007). In addition to the standard requirements for V to be valid as an instrument, we need V to be continuous and have large support, to serve as a so-called special regressor. To keep notation simple, we let

$$\Lambda \equiv \frac{DT}{\left(\gamma - \mathbb{E}\left(U_\gamma\right)\right) f_v\left(v\right)} Y, \quad \Pi \equiv \frac{DT}{\left(\gamma - \mathbb{E}\left(U_\gamma\right)\right) f_v\left(v\right)}, \tag{1.2.7}$$

where $T \equiv \mathbb{I}(-\gamma_0 \leq v_i \leq \gamma)$, γ_0 and γ are two positive numbers, and

$$U_\gamma = \begin{cases} \gamma & U > \gamma \\ U & -\gamma_0 \leq U \leq \gamma \\ -\gamma_0 & U < -\gamma_0 \end{cases}.$$

The following identification result is from Lewbel (2007).

Lemma 1.2.7 (**Identification**) *Under Assumption 8, 9 and 10, let* $p_D(v) \equiv \mathbb{E}(D|v)$, *then*

$$\mathbb{E}(\Lambda) = \mathbb{E}(Y^*) - \frac{cov(Y^*, U_\gamma)}{\gamma - \mathbb{E}(U_\gamma)}, \quad \mathbb{E}(\Pi) = 1$$

$$\text{var}(\Lambda) = \frac{1 + o(1)}{(\gamma - \mathbb{E}(U_\gamma))^2} \int_{-\gamma_0}^{\gamma} \frac{\mathbb{E}(Y^{*2} D|v)}{f_v(v)} dv \asymp \frac{1}{\gamma_n^2} \int_{-\gamma_0}^{\gamma_n} \frac{p_D(v)}{f_v(v)} dv.$$

From Lemma 1.2.7, $\mathbb{E}(Y^*)$ is identified by $\lim_{\gamma \to \infty} \Lambda$. To get a consistent estimate, we only need to let γ go to infinity while γ_0 can be a fixed number. The sample counterpart estimator is as follows:

$$\widehat{\mu}_n = \frac{\frac{1}{n}\sum_{i=1}^n \frac{D_i T_{ni}}{f_v(v_i)} Y_i}{\frac{1}{n}\sum_{i=1}^n \frac{D_i T_{ni}}{f_v(v_i)}}, \tag{1.2.8}$$

where $T_{ni} \equiv \mathbb{I}(-\gamma_0 \leq v_i \leq \gamma_n)$, $\gamma_n \to \infty$, as $n \to \infty$. We divide both the numerator and denominator by $\gamma_n - \mathbb{E}(U_n)$, then

$$\widehat{\mu}_n = \frac{\frac{1}{n}\sum_{i=1}^n \frac{D_i T_{ni}}{(\gamma_n - \mathbb{E}(U_n)) f_v(v_i)} Y_i}{\frac{1}{n}\sum_{i=1}^n \frac{D_i T_{ni}}{(\gamma_n - \mathbb{E}(U_n)) f_v(v_i)}}. \tag{1.2.9}$$

Similarly, we let Λ_{ni} and Π_{ni} denote the numerator and denominator in $\widehat{\mu}_n$ respectively. Since the denominator in equation (1.2.9) converges to one in probability and the struc-

tures of the denominator and numerator of $\widehat{\mu}_n$ are similar, we focus on the analysis on the numerator $\frac{1}{n}\sum_{i=1}^{n}\Lambda_{ni}$. The asymptotics for $\widehat{\mu}_n$ can be derived using the delta method.

If we drop γ_0 in T_{ni}, $var\left(\Lambda_{ni}\right)$ is possibly infinite for any γ_n. Thus γ_0 is necessary to be included. We denote the bias term and variance term for Λ_{ni} as

$$\mathcal{B}_n \equiv -\frac{cov\left(Y^*, U_n\right)}{\gamma_n - \mathbb{E}\left(U_n\right)}, \quad \sigma_n^2 \equiv \text{var}\left(\Lambda_{ni}\right).$$

1.2.4 Rates and Limiting Distribution

The following Lemma confirms the heavy-tail problem of estimator (1.2.8).

Lemma 1.2.8 *Under Assumption 3,* $\sigma_n^2 \to \infty$, *as* $\gamma_n \to \infty$.

We apply the Lindeberg-Feller CLT for $\frac{1}{n}\sum_{i=1}^{n}\Lambda_{ni}$. The Lindeberg condition for $\frac{1}{n}\sum_{i=1}^{n}\Lambda_{ni}$ is: for any $\varepsilon > 0$,

$$\lim_{n\to\infty}\mathbb{E}\left(\frac{\left(\Lambda_{ni} - \mathbb{E}\left(\Lambda_{ni}\right)\right)^2}{\sigma_n^2}\mathbb{I}\left[\frac{\left(\Lambda_{ni} - \mathbb{E}\left(\Lambda_{ni}\right)\right)^2}{\sigma_n^2} > n\varepsilon\right]\right) = 0. \tag{1.2.10}$$

We similarly let Ψ denote trimming parameters that satisfy the Lindeberg condition

$$\Psi = \bigcap_{\varepsilon>0}\left\{\gamma_n \left| \lim_{n\to\infty}\mathbb{E}\left(\frac{\left(\Lambda_{ni} - \mathbb{E}\left(\Lambda_{ni}\right)\right)^2}{\sigma_n^2}\mathbb{I}\left[\frac{\left(\Lambda_{ni} - \mathbb{E}\left(\Lambda_{ni}\right)\right)^2}{\sigma_n^2} > n\varepsilon\right]\right) = 0\right.\right\}.$$

It is straightforward to verify that Λ_{ni} satisfies all the assumptions given in the previous subsection. Applying Corollary 1.2.6, we get that the condition

$$nf_v^2\left(\gamma_n\right)\int_{-\gamma_0}^{\gamma_n}\frac{p_D\left(v\right)}{f_v\left(v\right)}dv \to \infty, \tag{1.2.11}$$

is a sufficient condition for the Lindeberg condition. By utilizing the specific structure of the estimator, the following theorem shows that condition (1.2.11) is necessary as well as sufficient for the Lindeberg condition, under a more general regularity condition (1.2.12) which further allows extremely thin-tailed densities as shown by Lemma 1.2.10.

Theorem 1.2.9 *Suppose Assumption 3, 35 ∼ 10 hold. Condition (1.2.11) is the sufficient condition to the Lindeberg condition (1.2.10). If there exists a differential function* $m(\gamma)$ *where* $0 < m(\gamma) < \gamma$ *such that*

$$\frac{f_v(\gamma - m(\gamma))}{f_v(\gamma)} = O(1) \qquad \text{and} \qquad \limsup_{\gamma\to\infty} \frac{(1 - m'(\gamma)) f_v(\gamma)}{f_v(\gamma - m(\gamma))} < 1. \tag{1.2.12}$$

Condition (1.2.11) is also the necessary condition.

Lemma 1.2.10 *If* $f_v(v)$ *decays at right tail the same order as* $\frac{1}{v^{1+c}}$, $v^{c_1}\exp(-v^{c_2})$, $\exp(-v^c)$, v^{-v^c}, $\exp(-\exp(v^c))$ *for any* $c, c_1, c_2 > 0$, *condition (1.2.12) holds.*

Based on the iff condition (1.2.11), in the next subsection, we introduce a condition that tells which world we are in and the optimal convergence condition for the trimming parameter γ_n to obtain the fastest possible convergence rate.

1.2.5 Nice or Ugly World and the Optimal Convergence Rate Condition

Under the iff condition

$$n f_v^2(\gamma_n) \int_{-\gamma_0}^{\gamma_n} \frac{p_D(v)}{f_v(v)} dv \to \infty, \tag{1.2.13}$$

the bias term may be the dominant term, depending on the tail thickness of $f_v(v)$. Note that

$$\sqrt{\frac{n}{\sigma_n^2}}\mathcal{B}_n \asymp \sqrt{\frac{nf_v^2(\gamma_n)\int_{-\gamma_0}^{\gamma_n}\frac{p_D(v)}{f_v(v)}dv}{f_v^2(\gamma_n)\left(\int_{-\gamma_0}^{\gamma_n}\frac{p_D(v)}{f_v(v)}dv\right)^2}}.$$

To have $\sqrt{\frac{n}{\sigma_n^2}}\mathcal{B}_n = O(1)$, we need $\frac{nf_v^2(\gamma_n)\int_{-\gamma_0}^{\gamma_n}\frac{p_D(v)}{f_v(v)}dv}{f_v^2(\gamma_n)\left(\int_{-\gamma_0}^{\gamma_n}\frac{p_D(v)}{f_v(v)}dv\right)^2} = O(1)$. Therefore, being in the nice world requires $f_v(\gamma)\int_{-\gamma_0}^{\gamma}\frac{p_D(v)}{f_v(v)}dv \to \infty$. Note that $p_D(v) \to 1$, as $v \to \infty$, so we only need

$$f_v(\gamma)\int_{-\gamma_0}^{\gamma}\frac{1}{f_v(v)}dv \to \infty, \text{ as } \gamma \to \infty. \tag{1.2.14}$$

Condition (1.2.14) is the condition that tells whether we are in the nice or the ugly world. If it holds, then we are in the nice world, otherwise we are in the ugly world, where either asymptotic normality fails or the bias term term dominates the distribution when asymptotic normality holds. We call equation (1.2.14) the tail condition.

The bias and variance tradeoff is similar to that for standard nonparametric estimation: the larger the γ_n, the smaller the bias is, but the larger the variance is. A closed form analytical expression for the minimizing RMSE is not available due to the complicated structure of the variance term. Nevertheless, the convergence rate of the estimator is either that of the bias or of the variance, whichever is slower. Therefore we get the fastest convergence rate by letting the bias term and variance term of the same magnitude, i.e., $\sqrt{\frac{n}{\sigma_n^2}}\mathcal{B}_n = 1$, which gives $\frac{1+o(1)}{(\gamma-\mathbb{E}(U_\gamma))^2}\int_{-\gamma_0}^{\gamma}\frac{\mathbb{E}\left(Y^{*2}D\middle|v\right)}{f_v(v)}dv = \frac{\mathrm{cov}(Y^*,U_\gamma)^2}{(\gamma-\mathbb{E}(U_\gamma))^2}$ and also implies asymptotic normality when we are in the nice world (based on the tail condition). Simplifying this equation and dropping small order terms, we have

$$\frac{1}{n}\int_{-\gamma_0}^{\gamma}\frac{\mathbb{E}\left(Y^{*2}D\middle|v\right)}{f_v(v)}dv = \mathrm{cov}(Y^*,U)^2. \tag{1.2.15}$$

The γ from this condition might not minimize RMSE, but it does minimize the rate of RMSE, so we call it the optimal convergence rate condition. We can estimate $\mathbb{E}\left(Y^{*2}D\middle|\, v\right)$ by regressing Y^2D on V. We then also need to estimate $\operatorname{cov}\left(Y^*,U\right)^2$, based on the model. After this, one could implement condition (1.2.15) to get γ. Note that this derivation based on the optimal convergence rate condition yields an optimal numerical value for γ, not just a rate.

To summarize, we first need to check the tail condition (1.2.14) to see if we are in the nice or the ugly world. If we are in the nice world, we choose a trimming parameter value based on the optimal convergence rate condition (1.2.15) to achieve the fastest convergence rate.

Below are the derivations of the convergence rate and which world we are in for our estimator with a given f_v. Note that when $f_v\left(v\right) \asymp \exp\left(-v\right)$, $f_v\left(\gamma\right)\int_{-\gamma_0}^{\gamma}\frac{p_D(v)}{f_v(v)}dv \asymp 1$. For specific density functions, $f_v\left(v\right) \asymp \exp\left(-v\right)$ is the boundary of the nice and the ugly world: for f_v with thicker tail, we are in the nice world; for f_v with thinner tail, we are in the ugly world.

Example 1: *If* $f_v\left(v\right) \asymp \frac{1}{v^{1+c}}$ *at the right tail for some* $c>0$, *then the optimal* γ_n *from condition (1.2.13) is that* $\gamma_n \asymp n^{\frac{1}{2+c}}$, and $\sqrt{\frac{n}{\sigma_n^2}}\mathcal{B}_n \asymp 1$, $\mathcal{B}_n \asymp n^{-\frac{1}{2+c}}$, $\sqrt{\frac{\sigma_n^2}{n}} \asymp n^{-\frac{1}{2+c}}$.

It is not hard to verify that the tail condition (1.2.14) holds here so we are in the nice world. Example 1 covers the case when V is distributed as a Cauchy or Student-t ($c \geq 1$).

Example 2: *If* $f_v\left(v\right) \asymp e^{-v^c}$ *at the right tail for some* $0<c<1$, *then the optimal* γ_n *from condition (1.2.13) is that* $\gamma_n \asymp \left(\log n\right)^{\frac{1}{c}}$, *and* $\sqrt{\frac{n}{\sigma_n^2}}\mathcal{B}_n \asymp 1$, $\mathcal{B}_n \asymp \left(\frac{1}{\log n}\right)^{\frac{1}{c}}$, $\sqrt{\frac{\sigma_n^2}{n}} \asymp \left(\frac{1}{\log n}\right)^{\frac{1}{c}}$.

In the case $f_v\left(v\right) \asymp e^{-v^c}$, $c<1$, the tail condition (1.2.14) holds, so we are in the nice world.

Example 3: *If* $f_v(v) \asymp e^{-v^c}$ *at the right tail for some* $c \geq 1$, *any* γ_n *from iff condition (1.2.13) will have* $\sqrt{\frac{n}{\sigma_n^2}}\mathcal{B}_n \to \infty$.

In the case $f_v(v) \asymp e^{-v^c}$, $c \geq 1$, the tail condition (1.2.14) fails. It is straightforward to verify that the dominant term is the bias term for any γ_n from the iff condition (1.2.13).

1.2.6 Asymptotic Normality and Inference

The following theorem is the main result in this paper. It states that when we are in the nice world (condition 1.2.14 holds), if we choose the trimming parameter using the optimal convergence rate condition (1.2.15), then we attain both the fastest possible convergence rate and asymptotic normality. If we are instead in the ugly world, then it may still be possible to make some progress by applying bias reduction techniques. We suggest using a Jackknife for this purpose. See Appendix 1.8 for details.

Theorem 1.2.11 *Let Assumption 3, 35, 8, 9, 10 hold. For* $f_v(v)$ *satisfies tail condition (1.2.14),* γ_n *from the optimal convergence rate condition (1.2.15), we have*

$$\sqrt{\frac{n}{\sigma_n^2}}\left[\frac{1}{n}\sum\nolimits_1^n \Lambda_{ni} - \mathbb{E}(Y^*) - \mathcal{B}_n\right] \xrightarrow{d} N(0,1),$$

where $\sqrt{\frac{n}{\sigma_n^2}}\mathcal{B}_n \asymp 1$ *and the convergence rate is the fastest.*

Proof.2 The conclusion follows immediately after the results in Section 1.2.5, by th Lindeberg-Feller central limit theorem. ■

The asymptotic distribution of estimator (1.2.8) follows immediately after Theorem 1.2.11.

Corollary 1.2.12 *Suppose all assumptions in Theorem 1.2.11 hold, we have*

$$\sqrt{\frac{n}{var\left[\Lambda_{ni}-\mathbb{E}\left(Y^{*}\right)\Pi_{ni}\right]}}\left[\widehat{\mu}_{n}-\mathbb{E}\left(Y^{*}\right)-\mathcal{B}_{n}\right]\overset{d}{\rightarrow}N\left(0,1\right),$$

where $\sqrt{\frac{n}{var[\Lambda_{ni}-\mathbb{E}(Y^{*})\Pi_{ni}]}}\mathcal{B}_{n}\asymp 1$ *and the convergence rate is the fastest.*

Proof.2 Not hard to see that the Lindeberg condition (1.2.10) also works for $\frac{1}{n}\sum_{1}^{n}\Pi_{ni}$. The rest of the proof is done by Theorem 1.2.11 and the delta method. ■

If we estimate the variance with $\widehat{\sigma}_{n}^{2}=\frac{1}{n}\sum_{i=1}^{n}\Lambda_{ni}^{2}-\left(\frac{1}{n}\sum_{i=1}^{n}\Lambda_{ni}\right)^{2}$, the following Lemma shows that $\frac{\widehat{\sigma}_{n}^{2}}{\sigma_{n}^{2}}\overset{p}{\rightarrow}1$, so we can estimate σ_{n}^{2} with above formula. This result is not trivial because $\sigma_{n}^{2}\rightarrow\infty$ as $n\rightarrow\infty$ here.

Lemma 1.2.13 *Let all assumptions in Theorem 1.2.11 hold, then we have* $\frac{\widehat{\sigma}_{n}^{2}}{\sigma_{n}^{2}}\overset{p}{\rightarrow}1$.

In the next section, we turn to case when $f_{v}\left(v\right)$ is unknown and estimated nonparametrically.

1.3 Estimation with Unknown f

f_{v} is usually unknown. In this section, we discuss the case when f_{v} is estimated nonparametrically.

We consider here the modified estimator (1.2.8) with estimated $\widehat{f_{v}}$,

$$\widehat{\mu}_{n}=\frac{\frac{1}{n}\sum_{i=1}^{n}\frac{D_{i}T_{ni}}{(\gamma_{n}-\mathbb{E}(U_{n}))\widehat{f}_{v}(v_{i})}Y_{i}}{\frac{1}{n}\sum_{i=1}^{n}\frac{D_{i}T_{ni}}{(\gamma_{n}-\mathbb{E}(U_{n}))\widehat{f}_{v}(v_{i})}}\equiv\frac{\frac{1}{n}\sum_{i=1}^{n}\widehat{\Lambda}_{ni}}{\frac{1}{n}\sum_{i=1}^{n}\widehat{\Pi}_{ni}},\tag{1.3.1}$$

where

$$\widehat{\Lambda}_{ni}\equiv\frac{D_{i}T_{ni}Y_{i}}{\left(\gamma_{n}-\mathbb{E}\left(U_{n}\right)\right)\widehat{f}_{v}\left(v_{i}\right)},\quad\widehat{\Pi}_{ni}\equiv\frac{D_{i}T_{ni}}{\left(\gamma_{n}-\mathbb{E}\left(U_{n}\right)\right)\widehat{f}_{v}\left(v_{i}\right)},$$

$$\widehat{f}_v(v_i) = \frac{1}{n-1}\sum_{j=1}^{n}\frac{1}{h}K\left(\frac{v_j - v_i}{h}\right).$$

$K(\cdot)$ is standard kernel function defined in Assumption 37.

The estimation of f_v introduces some new problems: the estimation of f_v is in expanding sets $[-\gamma_0, \gamma_n]$; the estimator now needs a linear representation. As shown in Wooldridge (2007), Hirano, Imbens, and Ridder (2003), Magnac and Maurin (2007), and many others, the estimator with estimated f_v can have smaller variance than the one using the true f_v. This is also the case here, however, the rate remains the same. For the convenience of inference, we prove the consistency of the bootstrap when we are in the nice world. Note that the convergence rate in this model is slower than root-n.

1.3.1 The Consistency of $\widehat{f}_v(v)$

To have a point-wise consistent estimate of $\widehat{f}_v(v)$, we need the number of observations around the point v to tend to infinity. We know that $f_v(\gamma_n) \asymp \inf_{v\in[-\gamma_0,\gamma_n]} f_v(v)$ for n large enough. So if $\widehat{f}_v(\gamma_n)$ is consistent for $f_v(\gamma_n)$, the point-wise consistency of $\widehat{f}_v(v)$ on the whole interval $[-\gamma_0, \gamma_n]$ should hold.

The standard nonparametric analysis (e.g., Li and Racine 2007) gives that

$$\mathbb{E}\left[\widehat{f}_v(v)\right] = f_v(v) + \frac{\kappa_q}{q!}f_v^{(q)}(v)\,h^q, \tag{1.3.2}$$

$$\operatorname{var}\left[\widehat{f}_v(v)\right] = \frac{\pi f_v(v)}{nh}, \tag{1.3.3}$$

where $\kappa_q \equiv \int v^q K(v)\,dv$, $\pi \equiv \int K(v)^2\,dv$, and q is the order of Kernel function K. From

equation (1.3.2) and (1.3.3),

$$\frac{\widehat{f}_v(v)}{f_v(v)} = 1 + \frac{\kappa_q}{q!}\frac{f_v^{(q)}(v)\,h^q}{f_v(v)} + O\left(\sqrt{\frac{\pi}{nhf_v(v)}}\right). \tag{1.3.4}$$

To control the variance term, we need the number of observations used to estimate $f_v(\gamma_n)$, $nhf(\gamma_n)$ to tend to infinity. The bias term could be controlled by using a high order kernel function with a bandwidth $h \asymp n^{-c}$, for some $c > 0$.

The optimal convergence rate condition (1.2.15) and the tail condition (1.2.14) imply that $nf_v(\gamma_n) \to \infty$. For the consistency of $\widehat{f}_v(v)$ on $[-\gamma_0, \gamma_n]$, we need a little bit stronger condition than that:

$$n^{1-c_h^*} f_v(\gamma_n) \to \infty, \tag{1.3.5}$$

for some $0 < c_h^* < 1$. The optimal convergence rate condition remains the same:

$$\frac{1}{n}\int_{-\gamma_0}^{\gamma_n} \frac{\mathbb{E}\left(Y^{*2}D\middle|\, v\right)}{f_v(v)}dv = \operatorname{cov}(Y^*, U)^2. \tag{1.3.6}$$

However, condition (1.3.5) and (1.3.6) place a more restrictive condition on $f_v(v)$:

$$\left(\int_{-\gamma_0}^{\gamma_n} \frac{1}{f_v(v)}dv\right)^{1-c_h^*} f_v(\gamma_n) \to \infty, \tag{1.3.7}$$

for some $0 < c_h^* < 1$. This is the new and stronger tail condition needed to be in the nice world with the estimated instead of true density. Condition (1.3.7) rules out $f_v(v) \asymp \exp(-v^c)$ for $c < 1$ in example 2. This is because the tail of that $f_v(v)$ is too thin to ensure the consistency of $\widehat{f}_v$ on the entire expanding sets, if we choose $h = n^{-c}$ for some $c > 0$.

Assumption 11 (Restriction on f_v) *For γ_n chosen from condition (1.3.6), $\frac{f_v(v+h)}{f_v(v)} =$*

$1+o(1)$, *for* $v \in [-\gamma_0, \gamma_n]$, *where* h *is the bandwidth used in the kernel function,* $h \asymp n^{-c}$, *for some* $c > 0$.

Assumption 11 is for the consistency of $\widehat{f}_v(v)$; intuitively, it says that the density of those observations used in estimation should be close to the density we estimate. It is not hard to verify that those f_v in Lemma 1.2.10 satisfy Assumption 11, so it is reasonable to impose this assumption.

Lemma 1.3.1 *For* γ_n *chosen from condition (1.3.6), under Assumption 11, if* $h \asymp n^{-c_h}$, *for some* $0 < c_h < 1$, *using Kernel defined in Assumption 37 with* $q > \frac{1-c_h}{c_h}$

$$\sup_{v \in [-\gamma_0, \gamma_n]} \left| \widehat{f}_v(v) - f_v(v) \right| = O\left(\left(\frac{\ln n}{nh} \right)^{\frac{1}{2}} \right).$$

Note that condition (1.3.6) can possibly give γ_n as fast as $n^{\frac{1}{2}}$, if the tail of $f_v(v)$ is thick enough. Hansen (2008) also obtains the uniform convergence rate of $\sup_{v \in [-\gamma_0, \gamma_n]} \left| \widehat{f}_v(v) - f_v(v) \right|$ on expanding set. However, this does not cover our result here, because our γ_n may go to infinity faster.

1.3.2 The First-Order Asymptotics

We consider the first-order asymptotics of $\frac{1}{n}\sum_1^n \widehat{\Lambda}_{ni}$. To simplify notation, let $m_{ni} \equiv \frac{D_i T_{ni} Y_i}{\gamma_n - \mathbb{E}(U_n)}$, then $\Lambda_{ni} \equiv \frac{m_{ni}}{f_v(v_i)}$, $\widehat{\Lambda}_{ni} \equiv \frac{m_{ni}}{\widehat{f}_v(v_i)}$.

Observe that

$$\widehat{\Lambda}_{ni} = \frac{m_{ni}}{\widehat{f}_v(v_i)} = \frac{m_{ni}}{f_v(v_i)} + \frac{m_{ni}\left(f_v(v_i) - \widehat{f}_v(v_i)\right)}{f_v^2(v_i)} + \frac{m_{ni}\left(f_v(v_i) - \widehat{f}_v(v_i)\right)^2}{f_v^2(v_i)\,\widehat{f}_v(v_i)}, \tag{1.3.8}$$

where the first two terms on the right hand side are the influence term and could be analyzed

using standard U-statistics, and the last term is the residual term, which is asymptotically negligible.

With the uniform convergence of $\widehat{f}_v(v)$ over the expanding sets, the following theorem gives the linear representation form, by applying the standard U-statistics (see Powell et al. 1989) technique on the influence term and showing the residual term is asymptotic negligible.

Theorem 1.3.2 *Suppose $f_v(v)$ satisfies condition (1.3.7). Let Assumption 3, 35 $\backsim$ 11, 37 hold. For γ_n chosen from condition (1.3.6), we set $h = n^{-c_h}$, $0 < c_h \leq c_h^*$, and $q > \frac{1-c_h}{c_h}$, then*

$$\frac{1}{n}\sum\nolimits_{i=1}^{n}\left(\widehat{\Lambda}_{ni} - \mathbb{E}(Y^*) - \mathcal{B}_n\right) = \frac{1}{n}\sum\nolimits_{i=1}^{n}\left(\Lambda_{ni} - \mathbb{E}(\Lambda_{ni}|v_i)\right) + o_p\left(\sqrt{\frac{\sigma_n^2}{n}}\right), \tag{1.3.9}$$

where the influence term is asymptotic normal and achieves the fastest rate of convergence, and $\sqrt{\frac{\sigma_n^2}{n}}\mathcal{B}_n \asymp 1$.

By Theorem 1.3.2 and for the same reason as in Corollary 1.2.12, we have the following Corollary.

Corollary 1.3.3 *Suppose all Assumptions in Theorem 1.3.2 hold, then*

$$\widehat{\mu}_n - \mathbb{E}(Y^*) - \mathcal{B}_n = \frac{1}{n}\sum\nolimits_{i=1}^{n}\left(\left[\Lambda_{ni} - \mathbb{E}(Y^*)\,\Pi_{ni}\right] - \mathbb{E}\left[\Lambda_{ni} - \mathbb{E}(Y^*)\,\Pi_{ni}|v_i\right]\right) + o_p\left(\sqrt{\frac{\sigma_n^2}{n}}\right), \tag{1.3.10}$$

where the influence term is asymptotic normal and achieves the fastest rate of convergence, and $\sqrt{\frac{\sigma_n^2}{n}}\mathcal{B}_n \asymp 1$.

Proof.2 It is not hard to see that the Lindeberg condition (1.2.10) also works for $\frac{1}{n}\sum_1^n \widehat{\Pi}_{ni}$.

The rest of the proof is done by Theorem 1.3.2 and the delta method. ■

The variance here is smaller than that in Corollary 1.2.12 with same degree of trimming, confirming previous results. However, the convergence rate remains the same.

1.3.3 Bootstrapping the Estimator

Suppose we have data $\{z_i\}_{i=1}^n$ and a statistic ϱ formed from $\{z_i\}_{i=1}^n$. The bootstrap randomly generates a series $\{z_i^*\}_{i=1}^n$ many times according to the empirical distribution of original series $\{z_i\}_{i=1}^n$, and then gets a new statistic ϱ^* based on $\{z_i^*\}_{i=1}^n$. ϱ^* is used to approximate the distribution of ϱ. The consistency of bootstrap has been discussed intensively in the literature. For an comprehensive review, see Horowitz (2001) and references therein.

Estimator (1.3.1) with a nonparametric estimated component is essentially a U-statistic. After some transformation, equation (1.3.8) becomes

$$\begin{aligned}\frac{1}{n}\sum\nolimits_{i=1}^{n}\widehat{\Lambda}_{ni} = \frac{1}{n}\sum\nolimits_{i=1}^{n}\frac{2m_{ni}}{f_v\left(v_i\right)} - \frac{1}{n\left(n-1\right)}\sum\nolimits_{i=1}^{n}\sum\nolimits_{j=1,j\neq i}^{n}Q_n\left(z_i,z_j\right) \\ + \frac{1}{n}\sum\nolimits_{i=1}^{n}\frac{m_{ni}\left(f_v\left(v_i\right)-\widehat{f}_v\left(v_i\right)\right)^2}{f_v^2\left(v_i\right)\widehat{f}_v\left(v_i\right)},\end{aligned} \tag{1.3.11}$$

where $Q_n\left(z_i,z_j\right) \equiv \frac{1}{2}\left(\frac{m_{ni}}{f_v^2(v_i)}+\frac{m_{nj}}{f_v^2(v_j)}\right)\frac{1}{h}K\left(\frac{v_j-v_i}{h}\right)$, Z denotes all the variables involved. The bootstrap for U-statistics is first discussed by Bickel and Freedman (1981), which gives conditions for the bootstrap to work. One condition is that second moment of $Q_n\left(z_i,z_j\right)$ is uniformly bounded which is not the case here. Chen, Linton, and Keilegom (2003) show bootstrap works for semiparametric estimates when the criterion function is not smooth but their results are in the regular case (root-n). So we need to show that the bootstrap works for estimator (1.3.1).

For notation we let variables with superscript $*$ be the newly generated variables from the empirical probability density function of $\{z_i\}_{i=1}^n$ with mass $\frac{1}{n}$ on each z_i, i.e., $\{z_i^*\}_{i=1}^n$ and $\widehat{\Lambda}_{ni}^*$ are the newly generated variables for $\{z_i\}_{i=1}^n$ and $\widehat{\Lambda}_{ni}$ respectively.

The theorem below says that the bootstrap technique indeed works for our estimator here, when we are in the nice world. The proof is tedious, but the idea of the proof is simple: we follow the standard proof of the consistency of the bootstrap for U-statistics while showing residual terms asymptotically negligible as in Section 1.3.2.

Theorem 1.3.4 *Under the same conditions in Theorem 1.3.2, and the bootstrap series* $\{z_i^*\}_{i=1}^n$ *are distributed as the empirical probability density function of* $\{z_i\}_{i=1}^n$ *with mass* $\frac{1}{n}$ *on each* z_i*, then*

$$\sqrt{\frac{n}{\mathbb{E}\left\{\left[\Lambda_{ni}-\mathbb{E}\left(\Lambda_{ni}|v_i\right)\right]^2\right\}}}\left[\frac{1}{n}\sum\nolimits_{i=1}^{n}\left(\widehat{\Lambda}_{ni}^*-\frac{1}{n}\sum\nolimits_{i=1}^{n}\widehat{\Lambda}_{ni}\right)\right]\xrightarrow{d}N\left(0,1\right).$$

1.4 Model with Additional Covariates

In this section, we generalize our identification and estimation to the case when we have additional covariates X. We first consider the case where we put no structural restrictions on how covariates affect the outcome, and then consider some parametric restrictions on the outcome equation. To simplify the already complicated analysis, we assume we know the joint density function of V and X. The results can readily but tediously be extended to the case with estimated density function, following the same line analysis as in section 1.3.

1.4.1 Nonparametric Estimates

The model is now as follows

$$Y = Y^* D,$$

$$D = \mathbb{I}\left(V - U \geq 0\right),$$

where we observe (Y, D, V, X), we do not observe U, and each variable is scalar except that X is $k \times 1$ vector. The object of interest is $\mathbb{E}\left(Y^*\right)$. For notational convenience, we let $Z = [Y, Y^*, D, V, U, X]$ denote all the variables involved here.

We basically maintain the previous assumptions, but now including X.

Assumption 12 *$V \perp U | X$, $\mathbb{E}\left(Y^* | U, X, V\right) = \mathbb{E}\left(Y^* | U, X\right)$, $0 < \underline{c} \leq var\left(Y^{*2} | U, X, V\right) \leq \mathbb{E}\left(Y^{*2} | U, X, V\right) \leq \overline{c} < \infty$, for any U, V, X.*

Assumption 13 *V is continuous with support $\mathbb{R}$. $\exists \gamma_0 > 0, \forall \gamma > 0, \inf\left\{f_v\left(v | x\right) | v \in [-\gamma_0, \gamma]\right\} > 0$.*

Assumption 14 *X lies in a compact set Ω_x, and $\inf_{x \in \Omega_x} f_x\left(x\right) > c_x > 0$. $f_v\left(v | x + h\right) = f_v\left(v | x\right)\left(1 + o(1)\right)$, for any $h = o\left(1\right)$.*

Assumption 15 (Restriction on $f(v|x)$) *$f_v\left(v | x\right)$ satisfies Assumption 3 at right tail for each fixed x.*

Assumption 13 is not restrictive, because we assume that X lies in a compact set.

Identification and the Estimator

For similar reasons as in Lemma 1.2.7,

$$\mathbb{E}\left(\frac{DY\mathbb{I}\left(-\gamma_0 \leq V \leq \gamma_n\left(X\right)\right)}{\left(\gamma_n\left(X\right)-\mathbb{E}\left(U_n|X\right)\right)f\left(V|\,X\right)}\middle|\,X\right)=\mathbb{E}\left(Y^*|\,X\right)-\frac{cov\left(Y^*,U_n|\,X\right)}{\gamma_n\left(X\right)-\mathbb{E}\left(U_n|X\right)}, \tag{1.4.1}$$

$$\mathbb{E}\left(\frac{D\mathbb{I}\left(-\gamma_0 \leq V \leq \gamma_n\left(X\right)\right)}{\left(\gamma_n\left(X\right)-\mathbb{E}\left(U_n|X\right)\right)f\left(V|\,X\right)}\middle|\,X\right)=1.$$

Then

$$\mathbb{E}\left[\frac{\mathbb{E}\left(\frac{DY\mathbb{I}(-\gamma_0\leq V\leq\gamma_n(X))}{(\gamma_n(X)-\mathbb{E}(U_n|X))f(V|X)}\middle|\,X\right)}{\mathbb{E}\left(\frac{D\mathbb{I}(-\gamma_0\leq V\leq\gamma_n(X))}{(\gamma_n(X)-\mathbb{E}(U_n|X))f(V|X)}\middle|\,X\right)}\right]=\mathbb{E}\left(Y^*\right)-\mathbb{E}\left[\frac{cov\left(Y^*,U_n|\,X\right)}{\gamma_n\left(X\right)-\mathbb{E}\left(U_n|X\right)}\right], \tag{1.4.2}$$

so $\mathbb{E}\left(Y^*\right)$ is identified by letting $\gamma_n\left(X\right)$ go to infinity for each X.

Let

$$\Lambda_{nj}^{(i)}\equiv\frac{D_jT_{nj}^{(i)}Y_j}{f_v\left(v_j|\,x_j\right)\left(\gamma_n\left(x_i\right)-\mathbb{E}\left(U_n|\,x_i\right)\right)},\quad \Pi_{nj}^{(i)}\equiv\frac{D_jT_{nj}^{(i)}}{f_v\left(v_j|\,x_j\right)\left(\gamma_n\left(x_i\right)-\mathbb{E}\left(U_n|\,x_i\right)\right)},$$

where $T_{nj}^{(i)}=\mathbb{I}\left(-\gamma_0\leq v_j\leq\gamma_n\left(x_i\right)\right)$, $\gamma_n\left(x_i\right)$ is the trimming index for x_i, then the sample counterpart estimator for equation (1.4.2) is:

$$\widehat{\widetilde{\mu}}_n=\frac{1}{n}\sum\nolimits_{i=1}^{n}\frac{\widehat{\mathbb{E}}\left(\Lambda_{ni}^{(i)}\middle|\,x_i\right)}{\widehat{\mathbb{E}}\left(\Pi_{ni}^{(i)}\middle|\,x_i\right)}=\frac{1}{n}\sum\nolimits_{i=1}^{n}\left[\frac{\frac{1}{n-1}\sum_{j\neq i}\Lambda_{nj}^{(i)}\frac{1}{h}K\left(\frac{x_j-x_i}{h}\right)}{\frac{1}{n-1}\sum_{j\neq i}\Pi_{nj}^{(i)}\frac{1}{h}K\left(\frac{x_j-x_i}{h}\right)}\right]. \tag{1.4.3}$$

Convergence Rate of the First-step Estimator and Lindeberg Condition.

After defining estimator (1.4.3), we treat x_i as a constant and discuss the first-step estimator

$$\frac{1}{n-1}\sum\nolimits_{j\neq i}\Lambda_{nj}^{(i)}\frac{1}{h}K\left(\frac{x_j-x_i}{h}\right),\quad \frac{1}{n-1}\sum\nolimits_{j\neq i}\Pi_{nj}^{(i)}\frac{1}{h}K\left(\frac{x_j-x_i}{h}\right). \tag{1.4.4}$$

We similarly define the bias and variance term for the first-step estimator:

$$\widetilde{\mathcal{B}}_n(x_i) \equiv -\frac{cov(Y^*, U_n|\, x_i)}{\gamma_n - \mathbb{E}(U_n|\, x_i)}, \quad \widetilde{\sigma}_n^2(x_i) \equiv \text{var}\left(\Lambda_{nj}^{(i)} \frac{1}{h} K\left(\frac{x_j - x_i}{h}\right)\right).$$

It is straightforward to check that

$$\widetilde{\sigma}_n^2(x_i) = \frac{(1+o(1))\int_{\mathbb{R}} K^2(u)\, du}{h(\gamma_n - \mathbb{E}(U_n|\, x_i))^2} \int_{-\gamma_0}^{\gamma_n(x_i)} \frac{\mathbb{E}\left(DY^{*2}\middle|\, x_i, v\right)}{f(v|\, x)} dv.$$

The variance of the first-step estimator is

$$\frac{n}{\widetilde{\sigma}_n^2(x_i)} \asymp h\frac{n}{\sigma_n^2},$$

which is slower than estimator (1.2.8) and (1.3.1).

Similar to Theorem 1.2.9, the following theorem shows that the iff condition of the Lindeberg condition for the first-step estimator.

Theorem 1.4.1 *Suppose Assumption 12 and 15 hold. If*

$$nf^2(\gamma_n(x_i)|\, x_i)\, h \int_{-\gamma_0}^{\gamma_n(x_i)} \frac{p_{D|x_i}(v)}{f(v|\, x)} dv \to \infty, \tag{1.4.5}$$

where $p_{D|x_i}(v) \equiv \mathbb{E}(D|\, V = v, X = x_i)$*. Then the Lindeberg condition of estimator (1.4.4) holds. If Lindeberg condition for estimator (1.4.4) holds and* $f(v|\, x)$ *satisfies similar regularity condition as in Theorem 1.2.9, then condition (1.4.5) holds.*

To be in the nice world, similar to Section 1.3.1, in this section, we need $f_{v|x}$ satisfy

$$\left(\int_{-\gamma_0}^{\gamma_n(x_i)} \frac{1}{f\left(v\middle| x\right)} dv\right)^{1-c_h^*} f_v\left(\gamma_n\left(x_i\right)\right) \to \infty, \tag{1.4.6}$$

for some $0 < c_h^* < 1$. Once we are in the nice world, similarly, we choose the trimming parameter from the following optimal convergence rate condition:

$$\frac{\int_{\mathbb{R}} K^2\left(u\right) du}{nh} \int_{-\gamma_0}^{\gamma_n(x_i)} \frac{\mathbb{E}\left(DY^2\middle| x_i, v\right)}{f\left(v\middle| x_i\right)} dv = \operatorname{cov}\left(Y^*, U_n\middle| x_i\right)^2, \tag{1.4.7}$$

which provide the trimming parameter that gives the fastest convergence rate for the first-step estimator.

Convergence Rate of the Plug-in Second-step Estimator

This portion of the analysis is standard and similar to the one in Section 1.3.2. Note that the structure of our estimator is essentially as follows:

$$\frac{\widehat{a}}{\widehat{b}} = -\frac{a}{b} + \frac{\widehat{a} - a}{b} - \frac{a\left(\widehat{b} - b\right)}{b^2} - \frac{\left(\widehat{a} - a\right)\left(\widehat{b} - b\right)}{b\widehat{b}} + \frac{a\left(\widehat{b} - b\right)^2}{b^2\widehat{b}}.$$

We first show the residual term (last two terms in above expression) is asymptotic negligible, then apply standard U-statistics technique on the influence term (first three terms in above expression).

Lemma 1.10.7 in the Appendix shows that the residual term is asymptotically negligible if we choose the trimming parameter using condition (1.4.7). After showing that the residual term is asymptotically negligible, we are able to give the first-order asymptotics of estimator (1.4.3). Since the proof on the influence term is standard and similar to the one in Lewbel

and Yang (2013), it is omitted.

Theorem 1.4.2 *Let all the Assumptions in Lemma 1.10.7 hold here, then*

$$\widehat{\widehat{\mu}}_n - \mathbb{E}(Y^*) - \mathbb{E}\left(\widetilde{\mathcal{B}}_n(x_i)\right)$$
$$= \frac{1}{n}\sum\nolimits_{i=1}^{n}\left(\frac{\Lambda_{ni}^{(i)}}{\mathbb{E}\left(\Pi_{ni}^{(i)}|x_i\right)} - \frac{\Pi_{ni}^{(i)}\mathbb{E}\left(\Lambda_{ni}^{(i)}|x_i\right)}{\left[\mathbb{E}\left(\Pi_{ni}^{(i)}|x_i\right)\right]^2} + \frac{\mathbb{E}\left(\Lambda_{ni}^{(i)}|x_i\right)}{\mathbb{E}\left(\Pi_{ni}^{(i)}|x_i\right)} - \mathbb{E}\left[\frac{\mathbb{E}\left(\Lambda_{ni}^{(i)}|x_i\right)}{\mathbb{E}\left(\Pi_{ni}^{(i)}|x_i\right)}\right]\right) + o_p\left(\sqrt{\frac{\sigma_n^2}{n}}\right),$$

where the influence term is asymptotic normal and achieves the fastest convergence rate, and $\sqrt{\frac{\sigma_n^2}{n}}\mathbb{E}\left(\widetilde{\mathcal{B}}_n(x_i)\right) \asymp 1$

The convergence rate of the two-step estimator here is the same as in Theorem 1.3.2. The slower convergence rate from the first-step estimator is smoothed out during the second-step estimation.

1.4.2 Semiparametric Estimates

The Model

We consider now the following semiparametric model,

$$Y = \left(X'\theta + e\right)D$$

$$D = \mathbb{I}(V - U \geqslant 0).$$

where we observe (Y, D, V, X), we do not observe U, and each variable is scalar except that X is $k \times 1$ vector. The object of interest is the parameter θ, a $k \times 1$ vector. We assume the moment condition that $\mathbb{E}(e|X) = 0$, so the only source of endogeneity comes from selection D.

We maintain assumptions 13∼15 here, and we modify Assumption 12 to accommodate current model for identification.

Assumption 16 *$V \perp U|\, X, \mathbb{E}\left(e|\, X\right) = 0, \mathbb{E}\left(XX'\right)$ is full rank, and $0 < \underline{c} \leq var\left(Y^{*2}|U, X, V\right) \leq \mathbb{E}\left(Y^{*2}|U, X, V\right) \leq \overline{c} < \infty$, for any U, V, X.*

The Estimator

Following the last subsection, we have

$$\mathbb{E}\left(\frac{DX\left(Y - X'\theta\right)\mathbb{I}\left(-\gamma_0 \leq V \leq \gamma\left(X\right)\right)}{\left(\gamma\left(X\right) - \mathbb{E}\left(U_\gamma|X\right)\right)f\left(V|\, X\right)}\right) = -\mathbb{E}\left[\frac{cov\left(Xe, U\right)}{\gamma\left(X\right) - \mathbb{E}\left(U_\gamma|X\right)}\right], \tag{1.4.8}$$

$$\mathbb{E}\left(\frac{D\mathbb{I}\left(-\gamma_0 \leq V \leq \gamma\left(X\right)\right)}{\left(\gamma\left(X\right) - \mathbb{E}\left(U_\gamma|X\right)\right)f\left(V|\, X\right)}\middle|\, X\right) = 1. \tag{1.4.9}$$

Therefore, we can identify θ by

$$\theta = \lim_{\gamma(X)\to\infty} \mathbb{E}\left(\frac{DXX'\mathbb{I}\left(-\gamma_0 \leq V \leq \gamma\left(X\right)\right)}{\left(\gamma\left(X\right) - \mathbb{E}\left(U_\gamma|X\right)\right)f\left(V|\, X\right)}\right)^{-1} \mathbb{E}\left(\frac{DXY\mathbb{I}\left(-\gamma_0 \leq V \leq \gamma\left(X\right)\right)}{\left(\gamma\left(X\right) - \mathbb{E}\left(U_\gamma|X\right)\right)f\left(V|\, X\right)}\right). \tag{1.4.10}$$

The sample counterpart estimator can be

$$\widehat{\theta} = \left(\frac{1}{n}\sum_{i=1}^{n} \frac{D_i T_{ni}^{(i)}}{\left(\gamma_n\left(x_i\right) - \mathbb{E}\left(U_\gamma|x_i\right)\right)f\left(v_i|\, x_i\right)} x_i x_i'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} \frac{D_i T_{ni}^{(i)}}{\left(\gamma_n\left(x_i\right) - \mathbb{E}\left(U_\gamma|x_i\right)\right)f\left(v_i|\, x_i\right)} x_i y_i\right). \tag{1.4.11}$$

For the asymptotic normality of each element in θ, the iff condition for the Lindeberg condition is the same for each element, since the variance of each component is of the same structure. Further, one can check that this iff condition is

$$nf^2\left(\gamma_n\left(x_i\right)|\, x_i\right) \int_{-\gamma_0}^{\gamma_n(x_i)} \frac{p_{D|x_i}\left(v\right)}{f\left(v|\, x_i\right)} dv \to \infty, \tag{1.4.12}$$

which is equation (1.4.5) without h.

Since θ is a vector, we choose the optimal convergence rate trimming parameter by minimizing the weighted RMSE of each component in θ by a non-negative weighting vector $\vartheta = (\vartheta_1, ..., \vartheta_k)$. The choice of ϑ affects the condition for choosing γ_n by a small amount. Because of this small discrepancy, we suggest using $\vartheta = (1, 0, ..., 0)$, i.e., putting all weight on the first term. With this weighting vector, by some simple calculation, we get the optimal convergence rate condition:

$$\frac{1}{n}\int_{-\gamma_0}^{\gamma_n(x_i)} \frac{\mathbb{E}\left(De^2 \middle| x_i, v\right)}{f\left(v \middle| x_i\right)} dv = \operatorname{cov}\left(e, U_n \middle| x_i\right)^2. \tag{1.4.13}$$

Similarly, for valid inference, we restrict our attention to the nice world where the density function $f(v|x)$ satisfies the tail condition (1.4.6). Using the Cramer-Wold device, the final asymptotics of $\widehat{\theta}$ is straightforward given what we have before. To save space, these details are omitted.

A Practical Alternative

Nonparametric estimation of $f(V|X)$ may be problematic in applications where X is moderate or high dimensional. To bypass this difficulty, one may put more structure on the model. For example, following Dong and Lewbel (2014), we assume that V is a linear function of X:

$$V = X'\alpha + \eta, \quad \eta \perp X, U, \quad \mathbb{E}(e|X, \eta) = \mathbb{E}(e|X), \tag{1.4.14}$$

and η is the new one-dimensional special regressor. We can first get $\widehat{\alpha}$ by doing linear regression of V on X and let $\widehat{\eta} = V - X'\widehat{\alpha}$. Then f_η could be estimated using a one-

dimensional nonparametric kernel density estimator.

We can identify θ by

$$\begin{aligned}\theta &= \lim_{\gamma\to\infty}\mathbb{E}\left(\frac{DXX'\mathbb{I}\left(-\gamma_0\leq V\leq\gamma\right)}{\left(\gamma-\mathbb{E}\left(U_\gamma\right)\right)f_\eta\left(\eta\right)}\right)^{-1}\mathbb{E}\left(\frac{DXY\mathbb{I}\left(-\gamma_0\leq V\leq\gamma\right)}{\left(\gamma-\mathbb{E}\left(U_\gamma\right)\right)f_\eta\left(\eta\right)}\right),\\ &= \lim_{\gamma\to\infty}\mathbb{E}\left(\frac{DXX'\mathbb{I}\left(-\gamma_0\leq V\leq\gamma\right)}{f_\eta\left(\eta\right)}\right)^{-1}\mathbb{E}\left(\frac{DXY\mathbb{I}\left(-\gamma_0\leq V\leq\gamma\right)}{f_\eta\left(\eta\right)}\right)\end{aligned} \tag{1.4.15}$$

where $\gamma-\mathbb{E}\left(U_\gamma\right)$ is canceled out in the second line. The iff condition for $\widehat{\theta}$ being asymptotically normal is equation (1.2.11), replacing v with η. To choose the trimming parameter, similar to the last subsection, we can let the weighting vector be $\vartheta=(1,0,...,0)$. Similarly, with this ϑ, the optimal convergence rate condition for choosing γ_n is

$$\frac{1}{n}\int_{-\gamma_0}^{\gamma_n}\frac{\mathbb{E}\left(De^2\middle|\,\eta\right)}{f_\eta\left(\eta\right)}d\eta=\operatorname{cov}\left(e,U_n\right)^2. \tag{1.4.16}$$

To be in the nice world, we need to restrict f_η to satisfy equation (1.3.7) (v replaced by η).

The sample counterpart estimator with $\widehat{f}$ can be

$$\widehat{\theta}=\left(\frac{1}{n}\sum_{i=1}^{n}\frac{D_iT_{ni}}{\widehat{f}_\eta\left(\widehat{\eta}_i\right)}x_ix_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{D_iT_{ni}}{\widehat{f}_\eta\left(\widehat{\eta}_i\right)}x_iy_i\right), \tag{1.4.17}$$

and $T_{ni}\equiv\mathbb{I}\left(-\gamma_0\leq\eta_i\leq\gamma_n\right)$. Because $\widehat{\eta}$ is root-n consistent, the asymptotics of $\widehat{\theta}$ will be not be affected by the preliminary estimation stage of η. Using the Cramer-Wold device, the final asymptotics of $\widehat{\theta}$ is a linear combination of the asymptotics in corollary 1.3.3 under almost the same conditions. Since the analysis is the same as before, to save space, we do not list formal results here.

1.5 Monte Carlo

To assess how our trimming criterion works in small samples, we provide two sets of Monte Carlo experiments. In the first experiment, all error terms are symmetric, while in the second experiment, there exists some asymmetry.

We set the number of observations to 200, 1000, and 5000, and the number of replications to 10000. We want to show that our way of trimming works well in small and in moderate large samples. We obtain five different estimates: one is our estimator with trimming parameter that is chosen from the optimal convergence rate condition (denoted as Full Trim in the table); two others are with halved and doubled that trimming parameter (denoted as Half Trim and Double Trim); the fourth one is the ordinary least square estimator without bias correction (denoted as OLS); last one is the Heckman's two-step estimator (denoted as Parametric). Heckman's estimator using MLE is more efficient than the two-step estimator when the error terms are normal. However, sometimes the Heckman MLE is hard to converge, so we choose the more robust two-step estimator as the benchmark.

The set up for the first Monte Carlo experiment is symmetric. We let e_1 and e_2 be the standard normal random variables, and set random variable V to the student-t distribution with the degree of freedom be 1, 3, or 4. The outcome and selection equation are:

$$Y = (1.5e_1 + 1.5e_2)\, D, \tag{1.5.1}$$

$$D = \mathbb{I}\left(V - 1.5e_2 \geq 0\right). \tag{1.5.2}$$

The expectation of the true underlying Y^* is zero, but one would see spurious negative

effects without any bias correction. We obtain two sets of estimates: one is obtained with true f_v, and the other one is obtained with $\widehat{f}_v$ estimated by the Gaussian kernel with bandwidth from the Silverman's Rule of Thumb. Note that the student-t distribution with 1 degree of freedom does not have finite variance, so for the latter case, we do not obtain estimates when V is t(1).

The set up for the second Monte Carlo experiment is asymmetric. We let e_3 be the standard uniform and e_4 is distributed as t(4), with V be the same as in the first experiment.

Again, e_3, e_4, V are independent. The outcome equation and selection equation are as follows:

$$Y = [2e_3 - 2\,|e_4|]\,D, \tag{1.5.3}$$

$$D = \mathbb{I}\,(V - 4e_3 \geq 0)\,. \tag{1.5.4}$$

The expectation of the true underlying Y^* is 0, and Y is asymmetric. Just as in the first experiment, we consider two different estimates: one with known f_v and one with estimated $\widehat{f}_v$. For the same reason as before, we do not consider the case with $\widehat{f}_v$ when V is distributed as t(1).

Note that Heckman's two-step estimator is consistent in the first experiment but is inconsistent in the second one because e_3 and e_4 are uniform.

All MC results are displayed in Table 3, 4 and Figure 1.2, 1.3 in the Appendix. We summarize our results in the tables as follows. First, the Heckman's two-step estimator is consistent in experiment one and inconsistent in experiment two. Our proposed estimator with the full trimming parameter performs reasonably well compared to Heckman's esti-

mator in experiment one. Our estimator performs similarly in both experiments. The OLS without bias correction shows large bias. Second, our estimator with the full trimming parameter outperforms those with the halved or doubled trimming parameter in terms of RMSE. The convergence rate of our estimator is indeed slower than root-n as seen from the RMSE in different sample sizes. One can also see that the heavier the tail that V has, the faster the convergence rate is, e.g., V being t(1) gives the fastest convergence rate. Last, the result with $\widehat{f}_v$ is similar to the result with true f_v.

Figure 1.2 and 1.3 show the results in both experiments when V distributed as t(3). The first three plots in both figures show the estimated probability density function (PDF) of our estimates with halved, full and doubled the optimized trimming parameter in different sample sizes. From the first plot, our estimator with halved trimming parameter shows big bias and the PDF of the estimates does not cover the true value zero. The third plot shows that our estimator with the doubled trimming parameter converges to zero very slowly. Our estimator with the optimized trimming parameter balances the bias and variance very well, as shown in the second plots. This is seen more clearly by the fourth plot, displaying our estimators with different degrees of trimming when the number of observations is 5000. The last two plots compare the PDF of our estimates with the full and doubled trimming parameter to the normal distribution with the same mean and variance as our estimates. They show that the PDF of our estimate with the full trimming is very close to normal while with the doubled trimming parameter, the PDF deviates from the normal distribution. The one in the second experiment even shows some degree of skewness. This shows that under trimming may lead to the failure of normality.

All in all, our MC results show that our estimator with the optimized trimming para-

meter works well in small and moderate large samples. Too small or too large trimming parameter gives undesirable RMSE. Under trimming may lead to the failure of asymptotic normality. MC results basically confirm previous theoretical findings.

1.6 Gender Wage Gap

1.6.1 Data

The gender wage gap problem fits right into the endogenous selection problem and has been studied extensively in the literature. Our analysis uses data from the Second Malaysian Family Life Survey (MFLS-2). This survey was conducted between August 1988 and January 1989 in Peninsular Malaysia. The MFLS-2 was developed by RAND and the National Population and Family Development Board of Malaysia. Previous work using this data set include Blau (1985, 1986), Vijverberg (1987), and Schafgans (1998, 2000). They found great discrepancies across different ethnicities in Malaysia. To simplify empirical analysis, we focus on the wage gap for a single ethnicity, specifically, between Malay men and women.

All monetary values are in 1985 prices, at an exchange rate of 2.48 Ringgit (M$). In line with similar studies, the exogenous variable we use is the non-employment income for individuals. The sources of non-employment income are unearned income (average yearly property income of the household in '00 M$'), house ownership (binomial 0 or 1), landholding (in '00 acres'), and other household members' yearly income (in '000 M$'). These wealth variables are assumed to affect individuals' reservation wage and hence thier decision to work, but are independent of the offered wage. We use minus the log of non-employment income as the special regressor V, so that V tending to positive infinity (no non-employment income at all) will force individuals to work. Summarized in Table 1.1,

Table 1.1: Summary Statistics

Variable Name	Male In	Male Out	Female In	Female Out
Hourly Wage	3.26	N.A.	2.00	N.A.
(M$)	(3.41)		(2.42)	
Age	30.17	22.61	28.67	26.80
(years)	(5.95)	(5.04)	(5.73)	(6.68)
Education Level	9.63	10.37	9.28	8.39
(years)	(3.36)	(3.11)	(3.91)	(4.00)
Unearned Income	5.56	10.21	7.42	11.96
('00M$)	(20.00)	(24.69)	(25.32)	(72.16)
Home Ownership	0.57	0.77	0.55	0.73
(0 or 1)	0.50	(0.42)	(0.50)	(0.44)
Land	0.66	0.92	0.59	0.97
('00acres)	(6.32)	(6.94)	(5.67)	(7.45)
Others' Income	3.99	6.30	7.31	7.38
('000M$)	(5.87)	(6.27)	(7.51)	(11.13)
Num of Observations	935	327	570	785

the non-employment income for individuals who are in the labor market is much worse than those who are not, which fits our intuition. Their introduction, however, does pose possible endogeneity problems arising from their dependence on previous earnings of the household. Following the literature, we only estimate over a young cohort 20-40, where non-employment income is more likely to be exogenous. The dependent variable is the log of the hourly wage rate. Other control variables are sex (0 denotes male and 1 denotes female), ages, squared ages, and education level (years of schooling). The notation for those variables is as follows: Y and Y^* are the observed log hourly wage rate and underlying log hourly wage rate respectively; X_V denote those non-employment income variables; d_S is the sex dummy; X_c are those control variables excluding sex dummy.

After dropping data with missing information, we have 2617 observations, including 1262 males and 1355 females. The participation rate of males is higher than that of females: 935 males but only 570 females are in the labor market. Mean values and standard errors (in parentheses) of those variables are summarized in Table 1.1. The first two columns and last

two columns are the statistics for males and females respectively, where "In" and "Out" denote inside and outside of the labor market respectively.

From Table 1.1, the hourly wage rate for females is only about 60% of that for males. Males not in the labor market are on average much younger than those in the labor market, while for females these two groups on average are about the same age. Males in the labor market on average are less educated than those who are not, but opposite is true for females.

We run the smoothed maximum score estimator to choose a weight β_V for the variables of non-employment income X_V to construct a special regressor $V = -\log(X_V'\beta_V)$. The maximum score estimator by Manski (1975, 1985) permits general forms of heteroskedasticity, but the convergence rate is cube-root-n and bootstrap is not consistent for inference. The smoothed maximum score estimator by Horowitz (1992) overcomes these issues; it converges faster and the bootstrap is consistent, so we use the smoothed maximum score technique here. We estimate the following model:

$$D = \mathbb{I}\left(\beta_0 - \log\left(X_V'\beta_V\right) + X_c'\beta_c - U \geqslant 0\right), \tag{1.6.1}$$

where $\beta_0, \beta_V, \beta_c$ are constant term, coefficients before X_V and X_c respectively. During estimation, we keep $\|\beta_V\|_2 = 1$, where $\|\cdot\|$ is the Euclidean norm. Following Horowitz (1992), we minimize

$$\frac{1}{n}\sum_{i=1}^{n}\Phi\left(\frac{\beta_0 - \log\left(X_V'\beta_V\right) + X_c'\beta_c}{h}\right),$$

where Φ is the cumulative density function (CDF) of the standard normal and $h = cn^{-\frac{1}{3}}$, c is some constant. We vary c from 0.5 to 1.5, and find that the results are not sensitive to the bandwidth we choose; we use the result from $c = 1$. The standard errors of our estimates

are obtained through the bootstrap with 200 replications. We finally get

$$V = -\log\left(\underset{(0.140)}{0.454} \times \text{unearned income} + \underset{(0.065)}{0.409} \times \text{others' income} + \underset{(0.089)}{0.316} \times \text{land} + \underset{(0.164)}{0.726} \times \text{houseown}\right),$$

where estimated coefficients with standard errors in parentheses are all positive significant as expected.

1.6.2 Oaxaca Decomposition

The gender wage gap can be decomposed into the part that is due to group differences in the magnitudes of the determinants, and the part that is due to group differences in the effects of those determinants. The latter difference is more reasonable to describe the wage gap than the original one, because the first difference can be explained by covariates. The Oaxaca decomposition (Oaxaca 1973) addresses this issue.

We illustrate the Oaxaca decomposition using our example. We further decompose (Y^*, X_c) into (Y_m^*, X_{mc}) and $\left(Y_f^*, X_{fc}\right)$ which are the variables for males and females respectively. Suppose the corresponding coefficients before X_m, X_f are θ_m, θ_f and let $\bar{\cdot}$ denote the average of the variable $\cdot$, then

$$\begin{aligned}
\overline{Y}_m^* - \overline{Y}_f^* &= \overline{X}_{mc}\theta_m - \overline{X}_{fc}\theta_f \\
&= \left(\overline{X}_{mc} - \overline{X}_{fc}\right)\theta_m + \overline{X}_{fc}\left(\theta_m - \theta_f\right) \qquad (1.6.2) \\
&= \left(\overline{X}_{mc} - \overline{X}_{fc}\right)\theta_f + \overline{X}_{mc}\left(\theta_m - \theta_f\right), \qquad (1.6.3)
\end{aligned}$$

where the first part in equation (1.6.2) and (1.6.3) is the gap attributed by endowment, and the second part is the gap by coefficients. Alternatively, we say the first part is explainable,

while the second part cannot be explained by what we observe. The second part is what we consider to be the gender wage gap.

Both equation (1.6.2) and (1.6.3) are possible decompositions. Without loss of generality, we adopt the one in equation (1.6.3).

1.6.3 Estimation

The outcome equation is model as follows:

$$Y^* = \theta_0 + X_c'\theta_c + d_s\theta_s + X_c' d_s \theta_{cs} + e, \tag{1.6.4}$$

$$Y = Y^* D \tag{1.6.5}$$

Relating these coefficients to the previous section, we have $\theta_m = \left[\theta_0, \theta_c'\right]'$, $\theta_f = \left[\theta_0 + \theta_s, (\theta_c + \theta_{cs})'\right]'$.

We find that estimation of $f(V|X)$ is sensitive to the bandwidth we choose, because the dimension of X is high. To make our results more robust, we impose more structure and adopt the simple approach in Section 1.4.2. We assume that

$$D = \mathbb{I}(\eta - U \geq 0),$$

where U is an unobservable, and η comes from

$$V = \alpha_0 + d_s\alpha_s + X_c'\alpha_c + \eta,$$

with assumption $\eta \perp X_c, d_s, U$, $\mathbb{E}(e|X, \eta) = E(e|X)$. In this way, we only need to run a linear regression to get $\widehat{\eta}$ and one dimension nonparametric estimation to get $\widehat{f}_\eta$. The final estimator is equation (1.4.17), with X being those regressors in equation (1.6.4).

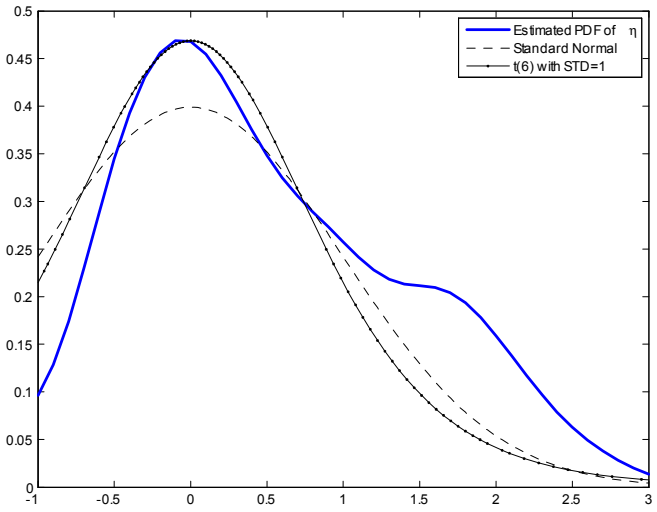

Figure 1.1: Estimated Density of η

To check whether we are in the nice world, Figure 1.1 displays the right tail of $\widehat{f_\eta}$ against a standard normal and a standardized student distribution with six degrees of freedom. From Figure 1.1, we argue that η has right tail behavior similar to that of t(6), indicating that we appear to be in the nice world.

We choose the trimming parameter γ_n from the optimal convergence rate condition (1.4.16). To estimate $\text{cov}(e, U)$, we impose the assumptions of Heckman's two-step estimator. However, note that these are assumptions used to get a good estimate of γ_n, they do not need to hold otherwise. We also provide estimates based on letting the trimming parameter equal $\gamma_n/2$ and $2\gamma_n$.

We compare our estimates with the ordinary least square estimator without any bias correction and with Heckman's two-step estimator, where the model is set as equation (1.6.1), (1.6.4), and (1.6.5). The estimated parameter with standard errors are displayed in Table 1.2. From Table 1.2, our estimates are sensitive to different trimming parameters, showing the importance of choosing it carefully as this paper provides. The coefficients of

Table 1.2: Estimation Results

	SR Full Trim	SR Double Trim	SR Half Trim	OLS	Parametric
Constant	-5.810^{***}	-3.965^{***}	-3.44^{***}	-4.466^{***}	-5.620^{***}
	(0.356)	(0.432)	(0.393)	(0.545)	(1.031)
Sex	2.82^{***}	1.294^{***}	0.440	1.226	0.907
	(0.565)	(0.474)	(0.563)	(0.824)	(0.863)
Age	0.338^{***}	0.200^{***}	0.140^{***}	0.257^{***}	0.318^{***}
	(0.024)	(0.030)	(0.029)	(0.038)	(0.058)
$\text{Age}^2/100$	-0.466^{***}	-0.243^{***}	-0.086^{*}	-0.352^{***}	-0.400^{***}
	(0.042)	(0.050)	(0.051)	(0.064)	(0.100)
Education	0.088^{***}	0.112^{***}	0.092^{***}	0.098^{***}	0.099^{***}
	(0.005)	(0.003)	(0.007)	(0.007)	(0.007)
Age×Sex	-0.226^{***}	-0.059^{*}	-0.025	-0.111^{*}	-0.092
	(0.040)	(0.033)	(0.041)	(0.058)	(0.060)
$\text{Age}^2/100\times\text{Sex}$	0.345^{***}	0.045	-0.041	0.141	0.100
	(0.070)	(0.057)	(0.071)	(0.098)	(0.100)
Education×Sex	0.037^{***}	-0.057^{***}	0.026^{***}	0.023^{**}	0.033^{***}
	(0.008)	(0.005)	(0.009)	(0.010)	(0.013)

Note: * significant at 10% level, ** significant at 5% level, *** significant at 1% level.

age and age^2 are expected to be positive and negative respectively, resulting in an inverted-U-shape type response centered around some positive value. All estimates of the coefficients before age and age^2 are as expected, except our estimator with halved the trimming parameter: the coefficient before age^2 is only significant at 10% level. For the coefficient before education×sex, our estimator with the doubled trimming parameter has an opposite sign than the others. For the parametric estimator, among the coefficients before the variables involved sex, only the one before education×sex is significant. Thus, if we do the Oaxaca decomposition and only keep those significant coefficients, the unexplained part will be negative, favoring women, which seems unlikely. To sum up, our estimator with the optimized amount of trimming delivers the most reasonable results.

The observed difference in the means of the log-wages for males and females ($\overline{Y}_m - \overline{Y}_f$) is 0.635. With OLS, the standard decomposition into the term $\left(\overline{X}_{mc} - \overline{X}_{fc}\right)\theta_m$ and $\overline{X}_{fc}\left(\theta_m - \theta_f\right)$ are 0.071 and 0.564 respectively. Only 11.2% of the difference in the mean

of log-wages can be explained by the superior endowment for males. For our estimator with the optimized trimming parameter, this percentage is 21.5%. From the parametric estimator, the percentage is 19.7%. After correcting for selection bias, the unexplained wage gap becomes smaller.

1.7 Conclusion

In this paper, we propose a general approach to trimming for heavy-tailed estimators. Unlike most of the previous literature, which either assumes the Lindeberg condition holds or imposes strong tail distribution assumptions, we instead find the largest range of possibly trimming parameter values that satisfy the Lindeberg condition without any tail distribution assumptions. We show a sharp distinction between a 'nice' and an 'ugly' world, which depends on details of the tail conditions. We demonstrate the results by working out the details for the special regressor estimator of endogenous selection models. A monte carlo experiment and an empirical study show that our approach works well in small samples.

The methods proposed here may be applied to a wide variety of other problems involving potentially heavy tailed estimators. Appendix 1.9 discusses some examples. Also, Appendix 1.8 shows that it may be possible to make progress using our approach even in the ugly world where standard inference is not possible.

Bibliography

[1] Abadie, A. (2003), Semiparametric Instrumental Variable Estimation of Treatment Response Models, Journal of Econometrics, 141, 777-806.

[2] Ahn, H., and J.L. Powell, (1993), Semiparametric Estimation of Censored Models with a Nonparametric Selection Mechanism, Journal of Econometrics, 58, 3-29.

[3] Andrews, D., and M. Schafgans (1998), Semiparametric Estimation of the Intercept of a Sample Selection Model, Review of Economic Studies, 65, 497–517.

[4] Bickel, P. (1982), On Adaptive Estimation, The Annals of Statistics, 10, 647-671.

[5] Bickel, P., and D. A. Freedman (1981), Some Asymptotic Theory for the Bootstrap, Annals of Statistics, Vol 9, 6, 1196-1217.

[6] Blau, D.M. (1985), The Effects of Economic Development on Life Cycle Wage Rates and Labor Supply Behavior in Malaysia, Journal of Development Economics, 19, 163-185.

[7] Blau, D.M., 1986. Self-employment, Earnings, and Mobility in Peninsular Malaysia, World Development, 14, 839–852.

[8] Brown, B. M. (1971), Martingale Central Limit Theorems, Annals of Mathematical Statistics, 42, 59-66.

[9] Campbell, J. and L. Hentschel (1992). No News is Good News: An Asymmetric Model of Changing Volatility in Stock Returns, Journal of Financial Economics 31, 281-313.

[10] Chaudhuri S., and J. B. Hill (2013), Robust Estimation for Average Treatment Effects, working paper.

[11] Chaudhuri, S., and H. Min (2012), Doubly-Robust Parametric Estimation in Moment Conditions Models with Missing Data, Mimeo.

[12] Chen, S. (1997), Semiparametric Estimation of Type-3 Tobit Model, Journal of Econometrics, 80, 1-34.

[13] Chen, X., O. Linton, and I. Van Keilegom, Estimation of Semiparametric Models When the Criterion Function is not Smooth, Econometrica, 71 (5), 1591-1608

[14] Chernozhukov, V., and I. Fernandez (2011). Inference for Extremal Conditional Quantile Models, with an Application to Birthweights, Review of Economic Studies, 78, 2, 559-589

[15] Csorgo, S., E. Haeusler, D.M., Mason (1988a), The Asymptotic Distribution of Trimmed Sums, Ann. Probab., 16, 672–699.

[16] Csorgo, S., E. Haeusler, D.M., Mason (1988b), A Probabilistic Approach to the Asymptotic Distribution of Sums of Independent, Identically Distributed Random Variables, Adv. Appl. Math., 9, 233–259.

[17] Crump, R., V. Hotz, G. Imbens, and O. Mitnik (2009), Dealing with Limited Overlap in Estimation of Average Treatment Effects, Biometrika, 96, 187–199.

[18] Das, M., W.K. Newey, F. Vella (2003), Nonparametric Estimation of Sample Selection Models, Review of Economic Studies 70, 33–58.

[19] Dong, Y., and A. Lewbel (2014), A Simple Estimator for Binary Choice Models with Endogenous Regressors, forthcoming, Econometrics Review.

[20] Embrechts, P., Kluppelberg, C. and Mikosch, T. (1997). Modelling Extremal Events for Insurance and Finance. Springer-Verlag: Frankfurt.

[21] Engle, R. amd V. Ng (1993). Measuring and Testing the Impact of News On Volatility, Journal of Finance 48, 1749-1778.

[22] Finkenstadt, B., and H. Rootzen (2003). Extreme Values in Finance, Telecommunications and the Environment. Chapman and Hall: New York.

[23] Frolich, M. (2004), Finite-Sample Properties of Propensity-Score Matching and Weighting Estimators, Review of Economics and Statistics, 86, 77–90.

[24] Hardle, W., and T. Stoker (1989), Investigating Smooth Multiple Regression by the Method of Average Derivatives, Journal of the American Statistical Association, Vol 84, 408, 986-995.

[25] Hahn, J. (1998), On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effect, Econometrica, 66, 315-331.

[26] Hansen, B. (2008), Uniform Convergence Rates for Kernel Estimation with Dependent Data, Econometric Theory, 24, 726-748.

[27] Heckman, J. (1976), The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models, Annals of Economic and Social Measurement, 5/4, 475-492.

[28] Heckman, J. (1979), Sample Selection Bias as a Specification Error, Econometrica, 47, 153-161.

[29] Heckman, J. (1990), Varieties of Selection Bias., American Economic Review: Papers and Proceedings, 313–318.

[30] Hill, B. (1975), A Simple General Approach to Inference about the Tail of a Distribution, Annals of Statistics, 3(5), 1163–1174.

[31] Hill, J.B., and E. Renault (2010), Generalized Method of Moments with Tail Trimming, working paper.

[32] Hill, J.B. and A. Shneyerov (2013), Are There Common Values on BC Timber Sales? A Tail-Index Nonparametric Test, Journal of Econometrics, 174, 144-164.

[33] Hirano, K., G. Imbens, and G. Ridder (2003), Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Scores, Econometrica, 71, 1161–1189.

[34] Honore, B.E., E. Kyriazidou, and C. Udry (1997), Estimation of Type-3 Tobit Models Using Symmetric Trimming and Pairwise Comparisons, Journal of Econometrics, 76, 107-128.

[35] Horowitz, J. L. (1992), A Smoothed Maximum Score Estimator for the Binary Response Model, Econometrica, 60, 3, 505-531.

[36] Horowitz, J. L. (2001), The Bootstrap in Econometrics, Handbook of Econometrics, Vol. 5, J.J. Heckman and E.E. Leamer, eds., Elsevier Science B.V., 2001, Ch. 52, pp. 3159-3228.

[37] Khan, S., and D. Nekipelov (2014), On Uniform Inference in Nonlinear Models with Endogeneity, working paper.

[38] Khan, S., and E. Tamer (2009), Irregular Identification, Support Conditions, and Inverse Weight Estimation, working paper.

[39] Khan, S., and E. Tamer (2010), Irregular Identification, Support Conditions, and Inverse Weight Estimation, Econometrica, 6, 2021-2042.

[40] Lee, L.F. (1994), Semiparametric Instrumental Variables Estimation of Simultaneous Equation Sample Selection Models, Journal of Econometrics, 63, 341–388.

[41] Lee, B., J. Lessler, and E. Stuart (2011), Weight Trimming and Propensity Score Weighting, PLOS One, 6.

[42] Lewbel, A. (1998), Semiparametric Latent Variable Model Estimation with Endogenous or Mismeasured Regressors, Econometrica, 66, 105-122.

[43] Lewbel, A. (2000), Semiparametric Qualitative Response Model Estimation with Unknown Heteroscedasticity and Instrumental Variables, Journal of Econometrics, 97, 145-177.

[44] Lewbel, A. (2007), Endogenous Selection or Treatment Model Estimation, Journal of Econometrics, 141, 777-806.

[45] Lewbel, A., and T. T. Yang (2013), Identifying the Average Treatment Effect in a Two Threshold Model, working paper.

[46] Li, Q., and J. S. Racine (2007), Nonparametric Econometrics: Theory and Practise, Princeton University Press.

[47] Li, Q., and J. Wooldridge (2002), Semiparametric Estimation of Partially Linear Models for Dependent Data with Generated Regressors, Econometric Theory, 18, 625-645.

[48] Magnac, T., and E. Maurin (2003), Identification and Information in Monotone Binary Models, Journal of Econometrics, 139, 76-104.

[49] Mandelbrot, B. (1963). The Variation of Certain Speculative Prices, Journal of Business 36, 394-419.

[50] Manski, C. (1975), Maximum Score Estimation of the Stochastic Utility Model of Choice, Journal of Econometrics, 3, 205-228

[51] Manski, C. (1984), Adaptive Estimation of Nonlinear Regression Models, Econometric Reviews, 3, 145-194

[52] Manski, C. (1985), Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator, Journal of Econometrics, 27, 313-334.

[53] Masry, E. (1996), Multivariate Local Polynomial Regression for Time Series: Uniform Strong Consistency Rates, Journal of Time Series Analysis, 17, 571-599.

[54] Oaxaca, R.L. (1973), Male–female Wage Differentials in Urban Labour Markets, International Economic Review 14, 693–709.

[55] Potter, F. (1993), The Effect of Weight Trimming on Nonlinear Survey Estimates, Proceedings of the Section on Survey Research Methods & Research. American Statistical Association.

[56] Powell, J.L. (1984), Least Absolute Deviations Estimation for the Censored Regression Model, Journal of Econometrics, 25, 305-325.

[57] Powell, J. L., J. H. Stock, and T. M. Stoker (1989), Semiparametric Estimation of Index Coefficients, Econometrica, 57, 1403-1430.

[58] Quenouille, M. H. (1949), Approximate Tests of Correlation in Time-Series, Journal of Royal Statistical Society, B, 11, 68-84.

[59] Resnick, S.I. (1997). Heavy Tail Modeling and Traffic Data, Annals of Statistics 25, 1805-1849.

[60] Robinson, P. M. (1988), Root N-Consistent Semiparametric Regression, Econometrica, 56, 931-954.

[61] Schafgans, M.M.A. (1998), Ethnic Wage Differences in Malaysia: Parametric and Semiparametric Estimation of the Chinese-Malay Wage-Gap, Journal of Applied Econometrics, 13, 481-504.

[62] Schafgans, M.M.A. (2000), Gender Wage Differences in Malaysia: Parametric and Semiparametric Estimation, 63, 351-378.

[63] Samorodnitsky, G., and M. S. Taqqu (1994), Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance, Chapman and Hall, New York.

[64] Serfling, R. J. (1980), Approximation Theorems of Mathematical Statistics, New York: Wiley.

[65] Silverman, B. W. (1978), Weak and Strong Uniform Consistency of the Kernel Estimate of a Density Function and its Derivatives, Annals of Statistics, 6, 177-184.

[66] Vella, F. (1992), Simple Tests for Sample Selection Bias in Censored and Discrete Choice Model, Journal of Applied Econometrics, 7, 413-421.

[67] Vella, F. (1998), Estimating Models with Sample Selection Bias: a Survey, Journal of Human Resources, 33, 127–169

[68] Vijverberg, P.W. (1987), Decomposing the Earnings Differentials in Peninsular Malaysia, Singapore Economic Review, 32, 24-36.

[69] Wooldridge, J.M. (1995), Selection Corrections for Panel Data Models under Conditional Mean Independence Assumptions, Journal of Econometrics, 68, 115–132.

[70] Wooldridge, J. M. (2007), Inverse Probability Weighted M-Estimation for General Missing Data Problems, Journal of Econometrics, 141, 1281-1301.

1.8 Appendix A: Jackknifing the Bias Term

The bias term here plays a very important role for estimator (1.2.8), causing trouble when it dominates the variance term. In this section, we propose a possible remedy for the bias term problem using Quenouille's (1949) jackknife estimator.

Let $\widehat{\theta}(z_1, z_2, ..., z_n)$ be a statistic over the whole sample with sample size n. Let $\mathbb{E}_n =$

$\mathbb{E}\left[\widehat{\theta}\left(z_1, z_2, ..., z_n\right)\right]$, and assume that

$$\mathbb{E}_n = \theta + \frac{c_1}{n} + \frac{c_2}{n^2} + ..., \tag{1.8.1}$$

where c_1, c_2 are some constants. Quenouille's method is based on sequentially deleting points z_i, and recomputing statistics $\widehat{\theta}_{(i)}$, then let

$$\widehat{\theta}_{(\cdot)} = \frac{1}{n}\sum\nolimits_{i=1}^{n} \widehat{\theta}_{(i)}.$$

The jackknife estimate is

$$\theta_J = n\widehat{\theta} - (n-1)\,\widehat{\theta}_{(\cdot)}. \tag{1.8.2}$$

Based on equation (1.8.1), it is not hard to see that $\mathbb{E}\left(\theta_J\right) = \theta + O\left(n^{-2}\right)$, so the bias term reduces from $O\left(n^{-1}\right)$ to $O\left(n^{-2}\right)$.

Back to our estimator here, we know from previous sections that

$$\mathbb{E}\left(\Lambda_{ni}\right) = \mathbb{E}\left(Y^*\right) - \frac{cov\left(Y^*, U_n\right)}{\gamma_n - \mathbb{E}\left(U_n\right)}, \quad \mathbb{E}\left(\Pi_{ni}\right) = 1.$$

The $\mathbb{E}\left(U_n\right)$ in the denominator is unknown and cannot be estimated. To apply jackknife technique, we modify the estimator a little bit

$$\phi_{ni} = \frac{D_i T_{ni} Y_i}{f_v\left(v_i\right)\gamma_n} \text{ and } \widehat{\phi}_n = \frac{1}{n}\sum\nolimits_{i=1}^{n} \phi_{ni}, \tag{1.8.3}$$

where $\widehat{\phi}_n$ is the modified estimator. Then by Lemma 1.2.7

$$\mathbb{E}\left(\widehat{\phi}_n\right) = \mathbb{E}\left(\phi_{ni}\right) = \mathbb{E}\left(Y^*\right) - \frac{\mathbb{E}\left(Y^* U_n\right)}{\gamma_n}. \tag{1.8.4}$$

Let the jackknifed estimator be

$$\widehat{\phi}_n^J = \frac{\gamma_n \widehat{\phi}_n - \gamma_{n-1}\left(\frac{1}{n}\sum_{i=1}^n \widehat{\phi}_n^{(i)}\right)}{\gamma_n - \gamma_{n-1}}, \tag{1.8.5}$$

where $\widehat{\phi}_n^{(i)}$ is the estimator in equation (1.8.3) dropping i-th observation with trimming parameter be γ_{n-1}. Then

$$\begin{aligned}
\mathbb{E}\left(\widehat{\phi}_n^J\right) &= \mathbb{E}\left(Y^*\right) - \frac{\mathbb{E}\left(Y^* U_n\right) - \mathbb{E}\left(Y^* U_{n-1}\right)}{\gamma_n - \gamma_{n-1}} \\
&= \mathbb{E}\left(Y^*\right) - \frac{1}{\gamma_n - \gamma_{n-1}} \int_{\gamma_{n-1}}^{\gamma_n} \mathbb{E}\left(Y^* \middle| u\right) u f_u\left(u\right) du.
\end{aligned} \tag{1.8.6}$$

Inspecting equation (1.8.6), the bias term is roughly $\gamma_n f_u\left(\gamma_n\right)$. Assumption 8 says that the second moment of U exists, which implies that

$$\lim_{u \to \infty} u^2 f_u\left(u\right) = 0.$$

Therefore $\gamma_n f_u\left(\gamma_n\right) = o\left(\frac{1}{\gamma_n}\right)$, thus the bias term is reduced. When $f_v\left(v\right)$ decays as exponential, for example, $f_v\left(v\right) = \exp\left(-v\right)$ at tails, then $\gamma_n = \log\left(n\right)$. U is usually thin tail than V, for simplicity let $f_u\left(u\right)$ be $\exp\left(-u\right)$ at tails as well. Then $\gamma_n f_u\left(\gamma_n\right) = \frac{\log(n)}{n}$, so the bias term is reduced dramatically in this case. In the case when $f_u\left(u\right) = 0$ between $\left[\gamma_{n-1}, \gamma_n\right]$, the bias term vanishes. From the above discussion, we know that jackknife works better in the case when both v and u are thin tails. More work needs to be done for the asymptotic

normality of the jackknifed estimator. We leave this for future research.

The advantage of the modified estimator (1.8.3) is that we can do the jackknife more easily and more efficiently. The disadvantage is that when U and Y^* are correlated, the original estimator is unbiased while the modified estimator is still biased. However, one usually knows a priori if endogeneity is likely. When the endogenous selection problem does exist, the modified estimator (1.8.3) is no worse than the original estimator.

1.9 Appendix B: Potential Extensions

In this section, we outline the way to do trimming for other important scenarios. These extensions are not trivial. In general, additional assumptions on the density function of certain unobservables are required. We point out where more regularity conditions are needed here and leave the details for future research.

1.9.1 The Average Derivative Estimator

Suppose we have $\mathbb{E}(Y|X) = m(X)$. Estimand is $\mu \equiv \mathbb{E}[m'(X)]$. Without loss of generality, we assume that X is a scalar. By $\mathbb{E}[m'(X)] = \mathbb{E}\left[-\frac{f'_x(X)}{f_x(X)}Y\right]$, Hardle and Stoker (1989) propose to estimate μ by $\frac{1}{n}\sum_{i=1}^{n}\frac{-f'_x(x_i)}{f_x(x_i)}y_i$, where $\{x_i, y_i\}_{i=1}^n$ are i.i.d. series from X, Y. To deal with the heavy-tails problem, we propose to estimate it by

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\frac{-f'_x(x_i)}{f_x(x_i)}y_i\mathbb{I}\left(-\gamma'_n \leq x_i \leq \gamma_n\right).$$

Under similar conditions to the previous sections, the iff condition to the Lindeberg condition for trimming parameters are equation (1.2.4) with v replaced by x.

By some derivations, we can obtain the bias term $\mathcal{B}_n$ and variance term σ_n^2 as follows:

$$\begin{aligned}\mathcal{B}_n &= -m(\gamma_n) f_x(\gamma_n) + m(-\gamma_n) f_x(-\gamma_n) - \int_{-\infty}^{-\gamma_n'} m'(x) f_x(x)\, dx - \int_{\gamma_n}^{\infty} m'(x) f_x(x)\, dx, \\ \sigma_n^2 &= (1+o(1)) \int_{-\gamma_n'}^{\gamma_n} \frac{f_x'(x)^2 \mathbb{E}\left(y^2 \middle| x\right)}{f_x(x)} dx.\end{aligned}$$

To get the optimal convergence rate condition as in equation (1.2.15) and the tail conditions with which we can apply standard inference as in equation (1.2.14), we can proceed as before, though doing so will require stronger regularity conditions on f_x.

1.9.2 The Special Regressor Estimator in Binary Choice model

Consider a standard threshold crossing binary choice model, where to simplify discussion, we assume regressors consist only of a constant μ and a single regressor V,

$$Y = \mathbb{I}(V + \mu - U \geq 0),$$

where V is a continuous variable that is independent of U and V has support on $\mathbb{R}$. Lewbel (2000) identifies the constant μ by $\mu = \mathbb{E}\left[\frac{Y - \mathbb{I}(V \geq 0)}{f_v(V)}\right]$, where f_v is the density function of V. The sample counterpart estimator is $\frac{1}{n}\sum_{i=1}^n \frac{y_i - \mathbb{I}(v_i \geq 0)}{f_v(v_i)}$, where $\{v_i, y_i\}_{i=1}^n$ are i.i.d. series. To deal with the heavy-tails problem, we propose to trim based on V,

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^n \frac{y_i - \mathbb{I}(v_i \geq 0)}{f_v(v_i)} \mathbb{I}\left(-\gamma_n' \leq v_i \leq \gamma_n\right).$$

Under similar conditions to previous sections, the iff condition to the Lindeberg condition for trimming parameters are equation (1.2.4).

By some derivations, we can have bias term $\mathcal{B}_n$ and variance term σ_n^2 as follows:

$$\begin{aligned}\mathcal{B}_n &= -\int_{-\infty}^{-\gamma_n'} F_u(\alpha+v)\,dv + \int_{\gamma_n}^{\infty}(1-F_u(\alpha+v))\,dv,\\ \sigma_n^2 &= (1+o(1))\int_{-\gamma_n'}^{\gamma_n}\frac{F_u(\alpha+v)(1-F_u(\alpha+v))}{f_v(v)}dv,\end{aligned}$$

where F_u is the cumulative density function of U. These expressions would form the basis of an anlysis to obtain the optimal convergence rate condition as in equation (1.2.15) and the tail conditions with which we can apply standard inference as in equation (1.2.14). Regularity conditions will need to be imposed on f_v and F_u.

1.9.3 The Propensity Score Weighted ATE Estimator

Consider the Average Treatment Effect (ATE) under unconfoundedness, where the treatment indicator D is independent of potential outcomes Y_1 and Y_0 after conditioning on control variables X. The observed outcome is $Y = Y_1 D + Y_0(1-D)$. To simplify analysis, we assume that X is a scalar variable here. The estimand ATE $\mu \equiv \mathbb{E}(Y_1 - Y_0) = \mathbb{E}\left[\frac{Y(D-P_x(X))}{P_x(X)(1-P_x(X))}\right]$, where P_x is the CDF of X. As discuss before, we propose to estimate μ by

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i(D_i - P_x(x_i))}{P_x(x_i)(1-P_x(x_i))}\mathbb{I}\left(-\gamma_n' \le x_i \le \gamma_n\right),$$

where $\{x_i, D_i, y_i\}_{i=1}^n$ are i.i.d. series from the model. The iff condition to the Lindeberg condition of $\widehat{\mu}$ could probably derived the same line analysis as in Section 1.2.2. Condition (1.2.5) is designed specifically for density function. Some similar conditions could probably be designed for cumulative density function.

The bias term and variance term for the term in $\widehat{\mu}$ is as follows:

$$\begin{aligned}\mathcal{B}_n &= -\int_{-\infty}^{-\gamma_n'} \left[\mathbb{E}\left(Y_1 \middle| x\right) - \mathbb{E}\left(Y_0 \middle| x\right)\right] f_x\left(x\right) dx - \int_{\gamma_n}^{\infty} \left[\mathbb{E}\left(Y_1 \middle| x\right) - \mathbb{E}\left(Y_0 \middle| x\right)\right] f_x\left(x\right) dx \\ \sigma_n^2 &= \int_{-\gamma_n'}^{\gamma_n} \left[\frac{\mathbb{E}\left(Y_1^2 \middle| x\right)}{P_x\left(x\right)} + \frac{\mathbb{E}\left(Y_0^2 \middle| x\right)}{1 - P_x\left(x\right)}\right] f_x\left(x\right) dx.\end{aligned}$$

As in the previous examples, these are the starting points for obtaining the optimal convergence rate condition as in equation (1.2.15) and the tail conditions with which we can apply standard inference as in equation (1.2.14), after imposing sufficient regularity conditions on f_x.

1.9.4 Heavy Tail Time Series Models

In a time series framework, many data are known to have heavy tails, but those heavy-tailed random terms are seldom independent. In many times series models, e.g., AR, ARCH, and GARCH, error terms are martingale difference sequences. To apply our approach, the Lindeberg condition we consider is now the one associated with a CLT for martingale difference arrays.

Suppose we have the martingale difference sequence $\{X_{k,T}\}_{k=1}^{T}$ with information set $\{\mathcal{F}_{k-1,T}\}_{k=1}^{T}$ for each $T > 0$. The following central limit theorem for martingale difference arrays is due to Brown (1971).

Theorem 1.9.1 (CLT for Martingale Difference Arrays) *If we have for any $\varepsilon > 0$,*

$$\frac{1}{T}\sum_{k=1}^{T} \mathbb{E}\left[X_{k,T}^2 \mathbb{I}\left(X_{k,T}^2 > \varepsilon T\right) \middle| \mathcal{F}_{k-1,T}\right] \xrightarrow{P} 0, \tag{1.9.1}$$

and $\frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\left[X_{k,T}^{2}\middle|\mathcal{F}_{k-1,T}\right]\xrightarrow{P}1$, *then we have* $\frac{1}{T}\sum_{k=1}^{T}X_{k,T}\xrightarrow{d}N(0,1)$.

Our iff condition to the Lindeberg condition (1.9.1), and the associated bias and variance with trimming, will differ from model to model. But the discussion of trimming parameters could follow the same line as before, with appropriate regularity conditions on error terms.

1.10 Appendix C: Some Technical Assumptions and Proof

Assumption 17 *The kernel functions* $K(v)$, $K(x)$ *have supports on* $[-1,1]$ *in* $\mathbb{R}$. *Each kernel function integrates to one over its support, is symmetric around zero, and has order* q, *i.e., for* $K(x)$,

$$\int_{\mathbb{R}}x^{l}K(x)dx=0\quad \textit{for } l<q,$$

$$\int_{\mathbb{R}}x^{q}K(x)dx\neq 0\ .$$

$\sup_{x\in[-1,1]}K(x)$ *is finite, and* $K(x)$ *satisfies Lipschitz condition, namely, there exists a* $c_K>0$, *such that*

$$\sup_{x\in[-1,1]}|K(x)|\leq c_K,\qquad |K(x+s)-K(x)|\leq c_K|s|.$$

This similarly holds for $K(v)$.

Lemma 1.10.1 *Under Assumption 3,* $\lim_{v\to\infty}vf_v(v)=0$.

Proof of Lemma 1.10.1.2 By Assumption 3 that $\frac{\widetilde{f}_v(v)}{f_v(v)}\asymp 1$, to prove $\lim_{v\to+\infty}vf_v(v)=0$, it is equivalent to show that

$$\lim_{v\to+\infty}v\widetilde{f}_v(v)=0.$$

We prove the conclusion by contradiction.

If not, there exists $c_1 > 0$, and a monotone increasing series going to infinity $\{\pi_n\}_{n=0}^{\infty}$, such that $\widetilde{f}_v(\pi_n) \geqslant \frac{c_1}{\pi_n}$. Let $\widetilde{f}_v(v) = \frac{c_1}{\pi_k}$, $v \in [\pi_{k-1}, \pi_k)$, then this $\widetilde{f}_v(v)$ will make $\int_{\pi_0}^{\pi_n} \widetilde{f}_v(v)\, dv$ the smallest among all $\widetilde{f}_v$ that satisfy monotonicity and $\widetilde{f}_v(\pi_n) \geqslant \frac{c_1}{\pi_n}$.

Fix π_0, π_n, easy to show that

$$\min_{\pi_1,\ldots,\pi_{n-1}} \int_{\pi_0}^{\pi_n} \widetilde{f}(v)\, dv = c_1 n \left(1 - \left(\frac{\pi_0}{\pi_n}\right)^{\frac{1}{n}}\right). \tag{1.10.1}$$

Since for any $c < 1$,

$$\lim_{x \to 0} \frac{1}{x}\left(1 - c^x\right) = -\ln c,$$

and by $\frac{\pi_0}{\pi_n} \to 0$, we have

$$\liminf_{n \to \infty} \int_{\pi_0}^{\pi_n} \widetilde{f}_v(v)\, dv \geq -\ln c,$$

for any $c < 1$. Therefore

$$\liminf_{n \to \infty} \int_{\pi_0}^{\pi_n} \widetilde{f}_v(v)\, dv \to \infty,$$

which contradicts with $\int_{\mathbb{R}} \widetilde{f}_v(v)\, dv \leq c_b < \infty$. ■

The intuition for Lemma 1.10.1 is as follows. Since $\int_1^{\infty} \frac{1}{x} dx = \infty$, $\int_1^{\infty} \frac{1}{x^{1+\varepsilon}} dx < \infty$, and $\int_{\mathbb{R}} f_v(v)\, dv = 1$, intuitively, $f_v(v)$ should decrease to 0 faster than $\frac{1}{v}$. Consequently, by Assumption 4, we have

$$f_v^2(\gamma) \int_0^{\gamma} \frac{\mathbb{E}\left(\varsigma^2 \middle| V = v\right)}{f_v(v)} dv \to 0, \text{ as } \gamma \to \infty, \text{or } \gamma \to -\infty. \tag{1.10.2}$$

Proof of Theorem 1.2.3.2 We first consider the case where one of the first two conditions in equation (1.2.4) hold. Without loss of generality, we assume that the first one holds.

Because $\mathbb{E}(x_{ni})$ is uniformly bounded and $\sigma_{ni}^2 \to \infty$, as $n \to \infty$, we have:

$$\Psi = \left\{ \gamma_n', \gamma_n \left| \lim_{n\to\infty} \mathbb{E}\left(\frac{x_{ni}^2}{\sigma_{ni}^2} \mathbb{I}\left[\frac{x_{ni}^2}{\sigma_{ni}^2} > n\varepsilon \right] \right) = 0 \right. \right\}. \tag{1.10.3}$$

For any fixed $\varepsilon > 0$, and γ_n', γ_n from the first condition in equation (1.2.4)

$$\mathbb{E}\left(\frac{x_{ni}^2}{\sigma_{ni}^2} \mathbb{I}\left[\frac{x_{ni}^2}{\sigma_{ni}^2} > n\varepsilon \right] \right) \leq \left(\int_{[-\gamma_n', 0]} + \int_{v \geq 0 \cap \left\{ v \left| \frac{x_{ni}^2}{\sigma_{ni}^2} > n\varepsilon \right. \right\}} \frac{\mathbb{E}\left(\varsigma^2 \middle| V = v \right)}{f_v(v)} dv \right) \Big/ \sigma_{ni}^2. \tag{1.10.4}$$

By the first condition in equation (1.2.4) we have

$$\left(\int_{[-\gamma_n', 0]} \frac{\mathbb{E}\left(\varsigma^2 \middle| V = v \right)}{f_v(v)} dv \right) \Big/ \sigma_{ni}^2 \to 0, \text{ as } n \to \infty. \tag{1.10.5}$$

So we only need to focus on the second integral in equation (1.10.4).

Consider the set when $v > 0$, for any fixed $\varepsilon > 0$, the following in equivalent

$$\left\{ v_i \left| \frac{x_{ni}^2}{\sigma_{ni}^2} > n\varepsilon, v > 0 \right. \right\}$$

$$\Leftrightarrow \left\{ v_i \left| \frac{\varsigma_i^2}{f_v^2(v_i)} \mathbb{I}(0 < v_i \leq \gamma_n) > n\varepsilon \int_{-\gamma_n'}^{\gamma_n} \frac{\mathbb{E}\left(\varsigma^2 \middle| V = v \right)}{f_v(v)} dv (1 + o(1)) \right. \right\}$$

$$\Leftrightarrow \left\{ v_i \left| \frac{\varsigma_i^2}{f_v^2(v_i)} \mathbb{I}(0 < v_i \leq \gamma_n) > n\varepsilon \int_{0}^{\gamma_n} \frac{\mathbb{E}\left(\varsigma^2 \middle| V = v \right)}{f_v(v)} dv (1 + o(1)) \right. \right\}$$

$$\Leftrightarrow \left\{ v_i \left| \frac{\varsigma_i^2}{\varepsilon} \mathbb{I}(0 < v_i \leq \gamma_n) > n f_v^2(v_i) \int_{0}^{\gamma_n} \frac{\mathbb{E}\left(\varsigma^2 \middle| V = v \right)}{f_v(v)} dv (1 + o(1)) \right. \right\} \tag{1.10.6}$$

where the third line holds by the first condition in equation (1.2.4). By assumption $f_v(v)$

decreases in order, condition

$$nf_v^2(\gamma_n)\int_0^{\gamma_n}\frac{\mathbb{E}\left(\varsigma^2\middle|V=v\right)}{f_v(v)}dv\to\infty$$

implies that the set in equation (1.10.6) is empty after some large n, so that we have

$$\int_{v\geq 0\cap\left\{v\left|\frac{x_{ni}^2}{\sigma_{ni}^2}>n\varepsilon\right.\right\}}\frac{\mathbb{E}\left(\varsigma^2\middle|V=v\right)}{f_v(v)}dv=0 \tag{1.10.7}$$

after some large n. Equation (1.10.4), (1.10.5), and (1.10.7) give that

$$\lim_{n\to\infty}\mathbb{E}\left(\frac{x_{ni}^2}{\sigma_{ni}^2}\mathbb{I}\left[\frac{x_{ni}^2}{\sigma_{ni}^2}>n\varepsilon\right]\right)=0,$$

which implies the Lindeberg condition.

For γ_n',γ_n from the third condition in equation (1.2.4), because $f_v(v)$ decreases in order at both tails, by the same logic as in the second part of the above proof, we have that $\left\{v_i\left|\frac{x_{ni}^2}{\sigma_{ni}^2}>n\varepsilon\right.\right\}$ is empty after some large n, for any fixed $\varepsilon>0$. This implies the Lindeberg condition. ■

Proof of Theorem 1.2.4.2 Without loss of generality, we consider the set

$$\mathbb{Q}\left(\gamma_n',\gamma_n\right)=\left\{v_i\left|nf_v^2(v_i)\int_0^{\gamma_n}\frac{\mathbb{E}\left(\varsigma^2\middle|V=v\right)}{f_v(v)}dv<\frac{c_1}{\varepsilon}\mathbb{I}\left(-\gamma_n'<v_i\leq\gamma_n\right)\right.\right\} \tag{1.10.8}$$

for some $c_1>0$, and fixed small $\varepsilon>0$, and the value

$$L\left(\gamma_n',\gamma_n\right)=\int_{v\in\mathbb{Q}(\gamma_n',\gamma_n)}\frac{\mathbb{E}\left(\varsigma^2\middle|V=v\right)}{f_v(v)}dv\Bigg/\int_{-\gamma_n'}^{\gamma_n}\frac{\mathbb{E}\left(\varsigma^2\middle|V=v\right)}{f_v(v)}dv \tag{1.10.9}$$

From the proof in Theorem 1.2.3, it is not hard to verify that set (1.10.8) is the essential part of set (1.2.2) and equation (1.10.9) is of the same order order as the expectation in the Lindeberg condition (1.2.10). Thus, the Lindeberg condition holds if and only if for all $\varepsilon > 0$

$$\lim_{n\to\infty} L\left(\gamma_n', \gamma_n\right) = 0. \tag{1.10.10}$$

Suppose we are in the first situation in condition (1.2.4), then

$$\int_{-\gamma_n'}^{\gamma_n} \frac{\mathbb{E}\left(\varsigma^2 \middle| V = v\right)}{f_v(v)} dv = \int_0^{\gamma_n} \frac{\mathbb{E}\left(\varsigma^2 \middle| V = v\right)}{f_v(v)} dv \left(1 + o(1)\right),$$

so we could disregard the effect γ_n' in this case. By equation (1.10.2), we can find $\{\gamma_n^*\}_{n=1}^{\infty}$, $\gamma_n^* \to \infty$, such that

$$n f_v^2(\gamma_n^*) \int_0^{\gamma_n^*} \frac{\mathbb{E}\left(\varsigma^2 \middle| V = v\right)}{f_v(v)} dv \asymp 1. \tag{1.10.11}$$

We prove the conclusion in two parts. In the first part, we prove that for any $\{\gamma_n\}_{n=1}^{\infty}$ with a sub-series going to infinity faster than γ_n^*, the Lindeberg condition fails to hold. In the second part, we prove that if a sub-series $\{\gamma_n^*\}_{n=1}^{\infty}$ could let the Lindeberg condition hold, then condition (1.2.11) can also give a corresponding sub-series $\{\gamma_n\}_{n=1}^{\infty}$ which is of the same order as $\{\gamma_n^*\}_{n=1}^{\infty}$. If any sub-series of those $\{\gamma_n^*\}_{n=1}^{\infty}$ fails the Lindeberg condition, our conclusion holds for sure. Combining the results above then implies the conclusion.

First, we set $\gamma_n = a_n \gamma_n^*$ as the trimming parameter, where $\{a_n\}_{n=1}^{\infty}$ is any series going to infinity. The results remain unchanged when only a sub-series of $\{a_n\}_{n=1}^{\infty}$ goes to infinity, for notational convenience, we say the whole series is the sub-series. By equation (1.10.2) and (1.10.11),

$$n f_v^2(a_n \gamma_n^*) \int_0^{a_n \gamma_n^*} \frac{\mathbb{E}\left(\varsigma^2 \middle| V = v\right)}{f_v(v)} dv = O(1).$$

By condition (1.2.5) in the Lemma and assumption 4 on $\mathbb{E}\left(\varsigma^2 \middle| V=v\right)$ we have,

$$\frac{f_v^2\left(\left(1-c_f\right)a_n\gamma_n^*\right)\int_0^{a_n\gamma_n^*}\frac{\mathbb{E}\left(\varsigma^2\middle|V=v\right)}{f_v(v)}dv}{f_v^2\left(a_n\gamma_n^*/\,a_f\right)\int_0^{a_n\gamma_n^*/a_f}\frac{\mathbb{E}(\varsigma^2|V=v)}{f_v(v)}dv}=O\left(1\right). \tag{1.10.12}$$

From equation (1.10.2) and $a_n\to\infty$, we know

$$\frac{nf_v^2\left(a_n\gamma_n^*/\,a_f\right)\int_0^{a_n\gamma_n^*/a_f}\frac{\mathbb{E}\left(\varsigma^2\middle|V=v\right)}{f_v(v)}dv}{nf_v^2\left(\gamma_n^*\right)\int_0^{\gamma_n^*}\frac{\mathbb{E}(\varsigma^2|V=v)}{f_v(v)}dv}=O\left(1\right). \tag{1.10.13}$$

So we have

$$nf_v^2\left(\left(1-c_f\right)a_n\gamma_n^*\right)\int_0^{a_n\gamma_n^*}\frac{\mathbb{E}\left(\varsigma^2\middle|\,V=v\right)}{f_v\left(v\right)}dv=O\left(nf_v^2\left(\gamma_n^*\right)\int_0^{\gamma_n^*}\frac{\mathbb{E}\left(\varsigma^2\middle|\,V=v\right)}{f_v\left(v\right)}dv\right)=O\left(1\right).$$

Therefore $\left[\left(1-c_f\right)a_n\gamma_n^*,a_n\gamma_n^*\right]\subseteq\mathbb{Q}\left(\gamma_n',a_n\gamma_n^*\right)$ for some small $\varepsilon>0$. Then after some large n

$$L\left(\gamma_n',a_n\gamma_n^*\right)\geq\frac{\int_{\left(1-c_f\right)a_n\gamma_n^*}^{a_n\gamma_n^*}\frac{\mathbb{E}\left(\varsigma^2\middle|V=v\right)}{f_v(v)}dv}{\int_0^{a_n\gamma_n^*}\frac{\mathbb{E}(\varsigma^2|V=v)}{f_v(v)}dv}\geq1-\frac{\int_0^{\left(1-c_f\right)a_n\gamma_n^*}\frac{\mathbb{E}\left(\varsigma^2\middle|V=v\right)}{f_v(v)}dv}{\int_0^{a_n\gamma_n^*}\frac{\mathbb{E}(\varsigma^2|V=v)}{f_v(v)}dv}.$$

So we have $\limsup_{n\to\infty}L\left(\gamma_n',a_n\gamma_n^*\right)>0$, by L'Hopital's rule and Assumption 6 on ω. Thus, equation (1.10.10) does not hold on series $\left\{\gamma_n',a_n\gamma_n^*\right\}_{n=1}^{\infty}$.

Second, we assume that a sub-series $\left\{\gamma_n^*\right\}_{n=1}^{\infty}$ will have the Lindeberg condition hold. For notational convenience, we say the sub-series is $\left\{\gamma_n^*\right\}_{n=1}^{\infty}$ itself. We set $\gamma_n=\gamma_n^*/a_f$. Then we say we must have

$$nf_v^2\left(\gamma_n^*/a_f\right)\int_0^{\gamma_n^*/a_f}\frac{\mathbb{E}\left(\varsigma^2\middle|\,V=v\right)}{f_v\left(v\right)}dv\to\infty.$$

Otherwise a sub series of $\left\{\gamma_n\right\}_{n=1}^{\infty}$ or $\left\{\gamma_n\right\}_{n=1}^{\infty}$ itself can have condition (1.10.11) hold. By

condition (1.2.5), set $a_n = 1$ in equation (1.10.12), and we can have that for that sub-series $[(1-c_f)\gamma_n^*, \gamma_n^*] \subseteq \mathbb{Q}(\gamma_n', \gamma_n^*)$, which leads to the contradiction that $\{\gamma_n^*\}_{n=1}^{\infty}$ fails the Lindeberg condition by the previous part of proof. Obviously, $\{\gamma_n\}_{n=1}^{\infty}$ is of the same order as $\{\gamma_n^*\}_{n=1}^{\infty}$.

We have proved that in the first situtation our conclusion holds. The proof our results hold in the second situation is the same as the previous one. In the third case, we define two sub series in the following way

$$\begin{aligned}
\{n_{i,+}\} &\equiv \left\{ n \left| \int_0^{\gamma_n} \frac{\mathbb{E}\left(\varsigma^2 \middle| V=v\right)}{f_v(v)} dv \geq \int_{-\gamma_n'}^{0} \frac{\mathbb{E}\left(\varsigma^2 \middle| V=v\right)}{f_v(v)} dv \right. \right\}, \\
\{n_{i,-}\} &\equiv \left\{ n \left| \int_0^{\gamma_n} \frac{\mathbb{E}\left(\varsigma^2 \middle| V=v\right)}{f_v(v)} dv < \int_{-\gamma_n'}^{0} \frac{\mathbb{E}\left(\varsigma^2 \middle| V=v\right)}{f_v(v)} dv \right. \right\}.
\end{aligned}$$

By the definition of the third situation in equation (1.2.4), both series have infinite elements. We then can apply the same analysis on both $\gamma_{n_{i,+}}$ and $\gamma'_{n_{i,-}}$, and the conclusion follows.

■

Proof of Lemma 1.2.5.2 We show the results one by one.

1. If $f_v(v) \asymp \frac{1}{v^{1+\delta}}$ at its right tail, by some simple calculations, condition (1.2.5) is satisfied.

2. Suppose $f_v(v) \asymp v^{c_1} \exp(-v^{c_2})$, for any $c_1 \geqslant 0, c_2 > 0$. Note that for any $a > 1$

$$\frac{f_v^2(v) \int_{-\gamma_0}^{a\gamma} \frac{1}{f_v(v)} dv}{f_v^2(\gamma) \int_{-\gamma_0}^{\gamma} \frac{1}{f_v(v)} dv} = \frac{f_v^2(v)}{f_v^{1+\varepsilon}(\gamma) f_v^{1+\varepsilon}(a\gamma)} \frac{f_v^{1+\varepsilon}(a\gamma) \int_{-\gamma_0}^{a\gamma} \frac{1}{f_v(v)} dv}{f_v^{1-\varepsilon}(\gamma) \int_{-\gamma_0}^{\gamma} \frac{1}{f_v(v)} dv}, \tag{1.10.14}$$

   By L'Hopital's rule, one can verify that $f_v^{1+\varepsilon}(a\gamma) \int_{-\gamma_0}^{a\gamma} \frac{1}{f_v(v)} dv \to 0$ and $f_v^{1-\varepsilon}(\gamma) \int_{-\gamma_0}^{\gamma} \frac{1}{f_v(v)} dv \to \infty$ for arbitrary small $\varepsilon > 0$. The deterministic component for the rate of $\frac{f_v^2(v)}{f_v^{1+\varepsilon}(\gamma) f_v^{1+\varepsilon}(a\gamma)}$

is the exponential term

$$\exp\left\{(1+\varepsilon)\left(1+a^{c_2}\right)\gamma^{c_2}-2v^{c_2}\right\}. \tag{1.10.15}$$

We can let $c_f=\frac{1}{2}\left(1-\frac{1}{a^{\frac{c_2}{2}}}\right)$ and it is then easy to check that $\frac{f_v^2\left(\left(1-c_f\right)a\gamma\right)}{f_v^{1+\varepsilon}(\gamma)f_v^{1+\varepsilon}(a\gamma)}=O\left(1\right)$ by verifying equation (1.10.15) with $v=\left(1-c_f\right)a\gamma$ and some small ε. Therefore, we have

$$\frac{f_v^2\left(\left(1-c_f\right)a\gamma\right)\int_{-\gamma_0}^{a\gamma}\frac{1}{f_v(v)}dv}{f_v^2\left(\gamma\right)\int_{-\gamma_0}^{\gamma}\frac{1}{f_v(v)}dv}=O\left(1\right),$$

which is the conclusion.

3. Suppose $f_v\left(v\right)\asymp v^{-v^c}$. We do the same transformation as equation (1.10.14). Then it is easy to verify that $f_v^{1+\varepsilon}\left(a\gamma\right)\int_{-\gamma_0}^{a\gamma}\frac{1}{f_v(v)}dv\to 0$ and $f_v^{1-\varepsilon}\left(\gamma\right)\int_{-\gamma_0}^{\gamma}\frac{1}{f_v(v)}dv\to\infty$ for arbitrary small $\varepsilon>0$. The deterministic component for the rate of $\frac{f_v^2(v)}{f_v^{1+\varepsilon}(\gamma)f_v^{1+\varepsilon}(a\gamma)}$ is $\gamma^{(1+\varepsilon)(1+a^c)\gamma^c}v^{-2v^c}$. By setting $c_f=\frac{1}{2}\left(1-\frac{1}{a^{\frac{c}{2}}}\right)$, for the same reason as before we have $\frac{f_v^2\left(\left(1-c_f\right)a\gamma\right)}{f_v^{1+\varepsilon}(\gamma)f_v^{1+\varepsilon}(a\gamma)}=O\left(1\right)$, and the desired result

$$\frac{f_v^2\left(\left(1-c_f\right)a\gamma\right)\int_{-\gamma_0}^{a\gamma}\frac{1}{f_v(v)}dv}{f_v^2\left(\gamma\right)\int_{-\gamma_0}^{\gamma}\frac{1}{f_v(v)}dv}=O\left(1\right).$$

4. If we have $f_v\left(v\right)\asymp\exp\left(-\exp\left(v^c\right)\right)$, we do the analysis in the opposite way. First verify that $f_v^{1-\varepsilon}\left(a\gamma\right)\int_{-\gamma_0}^{a\gamma}\frac{1}{f_v(v)}dv\to\infty$ and $f_v^{1+\varepsilon}\left(\gamma\right)\int_{-\gamma_0}^{\gamma}\frac{1}{f_v(v)}dv\to 0$ for arbitrary small $\varepsilon>0$. The deterministic component for the rate of $\frac{f_v^2(v)}{f_v^{1-\varepsilon}(\gamma)f_v^{1-\varepsilon}(a\gamma)}$ is

$$\exp\left\{\exp\left(\gamma^c+\log\left(1-\varepsilon\right)\right)+\exp\left(a^c\gamma^c+\log\left(1-\varepsilon\right)\right)-\exp\left(v^c+\log\left(2\right)\right)\right\}. \tag{1.10.16}$$

Then for any $0<c_f<1$, $v=\left(1-c_f\right)a\gamma$ will let equation (1.10.16) go to infinity which

is equivalent to $\frac{f_v^2(v)}{f_v^{1-\varepsilon}(\gamma) f_v^{1-\varepsilon}(a\gamma)} \to \infty$. Then we have the result, for any $0 < c_f < 1$,

$$\frac{f_v^2\left(\left(1-c_f\right)a\gamma\right)\int_{-\gamma_0}^{a\gamma}\frac{1}{f_v(v)}dv}{f_v^2\left(\gamma\right)\int_{-\gamma_0}^{\gamma}\frac{1}{f_v(v)}dv} \to \infty.$$

■

Proof of Lemma 1.2.7.2 First

$$\begin{aligned}
\mathbb{E}\left(\Lambda_{ni}\right) &= \frac{1}{\gamma_n - \mathbb{E}\left(U_n\right)}\mathbb{E}\left[\int_{\mathbb{R}}\frac{D\mathbb{E}\left(Y^*|u,v\right)T_{ni}}{f_v\left(v\right)}f\left(v|u\right)dv\right] \\
&= \frac{1}{\gamma_n - \mathbb{E}\left(U_n\right)}\mathbb{E}\left[\mathbb{E}\left(Y^*|u\right)\int_{\mathbb{R}}\mathbb{I}\left(v-u\geq 0\right)\mathbb{I}\left(-\gamma_0 \leq v \leq \gamma_n\right)dv\right] \\
&= \mathbb{E}\left(Y^*\right) + \frac{\mathbb{E}\left[\left(\mathbb{E}\left(Y^*|u\right) - \mathbb{E}\left(Y^*\right)\right)\left(\gamma_n - U_n\right)\right]}{\gamma_n - \mathbb{E}\left(U_n\right)} \\
&= \mathbb{E}\left(Y^*\right) - \frac{cov\left(Y^*, U_n\right)}{\gamma_n - \mathbb{E}\left(U_n\right)},
\end{aligned}$$

where by Assumption 8 the fourth line is finite. Similarly

$$\mathbb{E}\left(\Pi_{ni}\right) = 1.$$

Furthermore

$$\begin{aligned}
\mathbb{E}\left(\Lambda_{ni}^2\right) &= \int_{\mathbb{R}}\int_{\mathbb{R}}\frac{DT_{ni}\mathbb{E}\left(Y^{*2}|u,v\right)}{\gamma_n^2 f_v^2\left(v\right)}f_v\left(v\right)f_u\left(u\right)dvdu \\
&= \frac{1}{\left(\gamma_n - \mathbb{E}\left(U_n\right)\right)^2}\int_{-\gamma_0}^{\gamma_n}\frac{\mathbb{E}\left(Y^{*2}D\middle|\, v\right)}{f_v\left(v\right)}dv. \qquad (1.10.17)
\end{aligned}$$

From the first line of of equation (1.10.17), and by Assumption 9, we know that

$$\mathbb{E}\left(\Lambda_{ni}^2\right) \asymp \frac{1}{\gamma_n^2}\int_{-\gamma_0}^{\gamma_n}\frac{p_D\left(v\right)}{f_v\left(v\right)}dv \qquad (1.10.18)$$

Since

$$\mathbb{E}\left(\Lambda_{ni}\right) = O\left(1\right),$$

and under some mild condition by Lemma 1.2.8

$$\mathbb{E}\left(\Lambda_{ni}^{2}\right) \to \infty,$$

then

$$\begin{aligned} var\left(\Lambda_{ni}\right) &= \mathbb{E}\left(\Lambda_{ni}^{2}\right)\left(1+o\left(1\right)\right) \\ &= \frac{1+o\left(1\right)}{\left(\gamma_n - \mathbb{E}\left(U_n\right)\right)^2}\int_{-\gamma_0}^{\gamma_n}\frac{\mathbb{E}\left(Y^{*2}D\middle|\, v\right)}{f_v\left(v\right)}dv \asymp \frac{1}{\gamma_n^2}\int_{-\gamma_0}^{\gamma_n}\frac{p_D\left(v\right)}{f_v\left(v\right)}dv. \end{aligned}$$

■

Proof of Lemma 1.2.8.2 Note that

$$\lim_{\gamma_n\to\infty}\frac{\int_{-\gamma_0}^{\gamma_n}\frac{p_D(v)}{f_v(v)}dv}{\int_0^{\gamma_n}vdv} = \lim_{\gamma_n\to\infty}\frac{1}{\gamma_n f_v\left(\gamma_n\right)} = \infty, \tag{1.10.19}$$

where the first equality holds by L'Hopital's rule and second equality holds by Lemma 1.10.1. Equation (1.10.19) implies that

$$var\left(\Lambda_{ni}\right) \asymp \frac{1}{\gamma_n^2}\int_{-\gamma_0}^{\gamma_n}\frac{p_D\left(v\right)}{f_v\left(v\right)}dv \to \infty.$$

Replace $\left(1-c_f\right)a_f\gamma$ with $a_f\gamma - m\left(\gamma\right)$ in the proof of Theorem 1.2.4, we can get the desired result. ■

Proof of Theorem 1.2.9.2 The sufficiency of condition (1.2.11) holds obviously by pre-

vious results.

We define the set

$$\mathbb{Q}(\gamma_n) = \left\{ v \left| nf_v^2(v) \int_{-\gamma_0}^{\gamma_n} \frac{p_D(v)}{f_v(v)} dv < \frac{c_1}{\varepsilon} \mathbb{I}(-\gamma_0 < v \leq \gamma_n) \right. \right\} \tag{1.10.20}$$

for some $c_1 > 0$, some small $\varepsilon > 0$, and the value

$$L(\gamma_n) = \int_{v \in \mathbb{Q}(\gamma_n)} \frac{1}{f_v(v)} dv \Bigg/ \int_{-\gamma_0}^{\gamma_n} \frac{p_D(v)}{f_v(v)} dv. \tag{1.10.21}$$

By $p_D(v) \to 1$, as $v \to \infty$ and assumption 9, not hard to verify that equation (1.10.21) is of the same order order as the expectation in the Lindeberg condition.

Suppose we have the opposite of condition (1.2.11) that a sub-series of $nf_v^2(\gamma_n) \int_{-\gamma_0}^{\gamma_n} \frac{p_D(v)}{f_v(v)} dv$ is $O(1)$. For the convenience of notation, let the original series be the sub-series. Let $c_2 = \sup \left\{ nf_v^2(\gamma_n) \int_{-\gamma_0}^{\gamma_n} \frac{p_D(v)}{f_v(v)} dv \right\}$. Then if we have $\frac{f_v^2(\gamma_n)}{f_v^2(v)} > \frac{c_2}{c_1}\varepsilon$, we have $\frac{1}{nf_v^2(v) \int_{-\gamma_0}^{\gamma_n} \frac{p_D(v)}{f_v(v)} dv} > \frac{\varepsilon}{c_1}$.

For easier exposition, we strengthen the assumption that f_v decreases in order to the assumption that f_v is monotone decreasing. By condition $\frac{f_v(\gamma - m(\gamma))}{f_v(\gamma)} = O(1)$, we have $\frac{f_v^2(\gamma)}{f_v^2(\gamma - m(\gamma))} > \frac{c_2}{c_1}\varepsilon$ for some small $\varepsilon > 0$. Using the result in last paragraph, we have $\frac{1}{nf_v^2(\gamma_n - m(\gamma_n)) \int_{-\gamma_0}^{\gamma_n} \frac{p_D(v)}{f_v(v)} dv} > \frac{\varepsilon}{c_1}$. Therefore we have $[\gamma_n - m(\gamma_n), \gamma_n] \subset \mathbb{Q}(\gamma_n)$, implying that

$$L(\gamma_n) \geq \int_{\gamma_n - m(\gamma_n)}^{\gamma_n} \frac{1}{f_v(v)} dv \Bigg/ \int_{-\gamma_0}^{\gamma_n} \frac{p_D(v)}{f_v(v)} dv.$$

By L'Hopital's rule and condition (1.2.12), we have

$$\liminf_{\gamma_n \to \infty} L(\gamma_n) \geq 1 - \limsup_{\gamma_n \to \infty} \frac{(1 - m'(\gamma_n)) f_v(\gamma_n)}{f_v(\gamma_n - m(\gamma_n))} > 0,$$

which conflicts with the Lindeberg condition. ■

Proof of Lemma 1.2.10.2 We prove this by construction of $m(\gamma)$ for each density function. For easier exposition, we show the results for when the tails of f_v decay following exactly the same functions as those listed in the lemma.

$$\frac{f_v(\gamma - m(\gamma))}{f_v(\gamma)} = O(1) \qquad \textit{and} \qquad \limsup_{\gamma\to\infty} \frac{(1 - m'(\gamma)) f_v(\gamma)}{f_v(\gamma - m(\gamma))} < 1. \tag{1.10.22}$$

1. Suppose $f_v(v) = \frac{1}{v^{1+c}}$, after some large v. Let $m(\gamma) = (1 - c_f)\gamma$, where $0 < c_f < 1$. Then one can immediately verify that condition (1.2.12) holds.

2. Suppose $f_v(v) = v^{c_1} \exp(-v^{c_2})$, after some large v. Let $m(\gamma) = \gamma^{1-c_2}$. Then

$$\begin{aligned} \frac{f_v(\gamma - m(\gamma))}{f_v(\gamma)} &= \left(\frac{\gamma}{\gamma - \gamma^{1-c_2}}\right)^{c_1} e^{\left(\gamma - \gamma^{1-c_2}\right)^{c_2} - \gamma^{c_2}} = e^{-c_2} + o(1), \\ m'(\gamma) &= (1 - c_2)\gamma^{-c_2} = o(1), \end{aligned}$$

   by which condition (1.2.12) holds.

3. Suppose $f_v(v) = v^{-v^c}$, after some large v. Let $m(\gamma) = \frac{1}{c}\gamma^{1-c}\frac{\log 2}{\log \gamma}$. Then

$$\begin{aligned} \frac{f_v(\gamma - m(\gamma))}{f_v(\gamma)} &= \gamma^{\gamma^c - (\gamma - m(\gamma))^c} \left(\frac{\gamma - m(\gamma)}{\gamma}\right)^{-(\gamma - m(\gamma))^c} \\ &= \gamma^{\gamma^c\left[1 - \left(1 - \frac{m(\gamma)}{\gamma}\right)^c\right]} \left[\left(1 - \frac{m(\gamma)}{\gamma}\right)^{-\frac{\gamma}{m(\gamma)}}\right]^{\frac{m(\gamma)}{\gamma}(\gamma - m(\gamma))^c} \\ &= \gamma^{c\gamma^{c-1}m(\gamma) + o\left(\gamma^{c-1}m(\gamma)\right)} (e + o(1))^{o(1)} = 2 + o(1), \\ m'(\gamma) &= \frac{1 - c}{c}\gamma^{-c}\frac{\log 2}{\log \gamma} - \frac{1}{c}\gamma^{-c}\frac{\log 2}{(\log \gamma)^2} = o(1), \end{aligned}$$

   by which condition (1.2.12) holds.

4. Suppose $f_v(v) = e^{-e^{v^c}}$, after some large v. Let $m(\gamma) = \frac{1}{c}\gamma^{1-c}e^{-\gamma^c}$. Then

$$\begin{aligned}
\frac{f_v(\gamma - m(\gamma))}{f_v(\gamma)} &= e^{e^{\gamma^c}\left[1-e^{(\gamma-m(\gamma))^c-\gamma^c}\right]} = e^{e^{\gamma^c}\left[1-e^{-cm(\gamma)\gamma^{c-1}+o\left(m(\gamma)\gamma^{c-1}\right)}\right]} \\
&= e^{e^{\gamma^c}\left[cm(\gamma)\gamma^{c-1}+o\left(m(\gamma)\gamma^{c-1}\right)\right]} \\
&= e + o(1), \\
m'(\gamma) &= \frac{1-c}{c}\gamma^{-c}e^{-\gamma^c} - e^{-\gamma^c} = o(1),
\end{aligned}$$

by which condition (1.2.12) holds.

■

Proof of Lemma 1.2.13.2 By Lemma 1.2.7, we know that $\sigma_n^2 = \mathbb{E}\left(\Lambda_{ni}^2\right)(1 + o(1))$. By Lemma 1.2.8, $\sigma_n^2 \to \infty$. So we have

$$\frac{\widehat{\sigma}_n^2}{\sigma_n^2} - 1 = \frac{\frac{1}{n}\sum_{i=1}^n\left[\Lambda_{ni}^2 - \mathbb{E}\left(\Lambda_{ni}^2\right)\right]}{\mathbb{E}\left(\Lambda_{ni}^2\right)} + o_p(1). \tag{1.10.23}$$

To show the conclusion, we only need to show that $\frac{\frac{1}{n}\sum_{i=1}^n\left[\Lambda_{ni}^2-\mathbb{E}\left(\Lambda_{ni}^2\right)\right]}{\mathbb{E}\left(\Lambda_{ni}^2\right)} = o_p(1)$. To this end, we show the variance of this term is $o(1)$.

$$var\left(\frac{\frac{1}{n}\sum_{i=1}^n\left[\Lambda_{ni}^2 - \mathbb{E}\left(\Lambda_{ni}^2\right)\right]}{\mathbb{E}\left(\Lambda_{ni}^2\right)}\right) = \frac{\frac{1}{n}\mathbb{E}\left(\Lambda_{ni}^4\right)}{\mathbb{E}\left(\Lambda_{ni}^2\right)^2} + o(1) = \frac{\frac{1}{n}\int_{-\gamma_0}^{\gamma_n}\frac{\mathbb{E}\left(DY^{*4}\right)}{f_v^3(v)}dv}{\left(\int_{-\gamma_0}^{\gamma_n}\frac{\mathbb{E}(DY^{*2})}{f_v(v)}dv\right)^2} + o(1).$$

By condition (1.2.15), we have $\frac{1}{n}\int_{-\gamma_0}^{\gamma_n}\frac{\mathbb{E}\left(DY^{*2}\right)}{f_v(v)}dv \asymp 1$. By condition (1.2.15) and (1.2.14), we have $nf_v(\gamma_n) \to \infty$. Therefore

$$var\left(\frac{\frac{1}{n}\sum_{i=1}^n\left[\Lambda_{ni}^2 - \mathbb{E}\left(\Lambda_{ni}^2\right)\right]}{\mathbb{E}\left(\Lambda_{ni}^2\right)}\right) \asymp \frac{1}{n}\int_{-\gamma_0}^{\gamma_n}\frac{\mathbb{E}\left(DY^{*4}\right)}{f_v(v)}\frac{1}{\left(nf_v(v)\right)^2}dv \to 0.$$

The convergence comes from condition (1.2.15) and $nf_v(\gamma_n) \to \infty$. Then conclusion is obtained by Markov's inequality. ■

Proof of Lemma 1.3.1.2 The proof is a modification of Masry (1996) and Li and Racine (2007).

The object of interest is

$$\sup_{v\in[-\gamma_0,\gamma_n]} \left|\widehat{f}_v(v) - f_v(v)\right|. \tag{1.10.24}$$

Decompose equation (1.10.24):

$$\sup_{v\in[-\gamma_0,\gamma_n]} \left|\widehat{f}_v(v) - f_v(v)\right| \leq \underbrace{\sup_{v\in[-\gamma_0,\gamma_n]} \left|\widehat{f}_v(v) - \mathbb{E}\left(\widehat{f}_v(v)\right)\right|}_{P_1} + \underbrace{\sup_{v\in[-\gamma_0,\gamma_n]} \left|\mathbb{E}\left(\widehat{f}_v(v)\right) - f_v(v)\right|}_{P_2} \tag{1.10.25}$$

For P_2, use equation (1.3.2)

$$P_2 \leq c_1 h^q, \tag{1.10.26}$$

for some $c_1 > 0$.

The rest of proof focus on P_1. Since $[-\gamma_0, \gamma_n]$ is compact for fixed n, we can cover it by L_n intervals $\{\mathbb{I}_{n,k}\}_{k=1}^{L_n}$ with length $l_n = \frac{2\gamma_n}{L_n}$. Let $v_{n,k}$ be an inner point of $\mathbb{I}_{n,k}$. Then

$$P_1 = \max_{1\leq k\leq L_n} \sup_{v\in\mathbb{I}_{k,n}} \left|\widehat{f}_v(v) - \mathbb{E}\left(\widehat{f}_v(v)\right)\right| \leq \underbrace{\max_{1\leq k\leq L_n} \sup_{v\in\mathbb{I}_{k,n}} \left|\widehat{f}_v(v) - \mathbb{E}\left(\widehat{f}_v(v_{k,n})\right)\right|}_{P_{11}}$$
$$+ \underbrace{\max_{1\leq k\leq L_n} \left|\widehat{f}_v(v_{k,n}) - \mathbb{E}\left(\widehat{f}_v(v_{k,n})\right)\right|}_{P_{12}} + \underbrace{\max_{1\leq k\leq L_n} \sup_{v\in\mathbb{I}_{k,n}} \left|\mathbb{E}\left(\widehat{f}_v(v_{k,n})\right) - \mathbb{E}\left(\widehat{f}_v(v)\right)\right|}_{P_{13}}.$$

For P_{12},

$$P_{12} = \max_{1\leq k\leq L_n} \left|\sum\nolimits_{i=1}^{n} w_{ni}(v_{k,n})\right|,$$

where

$$w_{ni}\left(v_{k,n}\right)=\frac{K\left(\frac{v_{k,n}-v_i}{h}\right)-\mathbb{E}\left(K\left(\frac{v_{k,n}-v_i}{h}\right)\right)}{nh}.$$

Easy to get that $\mathbb{E}\left(w_{ni}^2\left(v_{k,n}\right)\right)\leq\frac{c_2}{n^2h}$, for some $c_2>0$. Then

$$p\left(P12>\eta\right)=p\left(\max_{1\leq k\leq L_n}\left|\sum_{i=1}^n w_{ni}\left(v_{k,n}\right)\right|>\eta\right)\leq L_n\sup_{v\in[-\gamma_0,\gamma_n]}p\left(\left|\sum_{i=1}^n w_{ni}\left(v\right)\right|>\eta\right). \tag{1.10.27}$$

Let

$$\lambda_n=\left(nh\ln n\right)^{\frac{1}{2}},$$

then by Assumption 37 and $c_h<1$,

$$\lambda_n\left|w_{ni}\left(v\right)\right|\leq c_K\left(\frac{\ln n}{nh}\right)^{\frac{1}{2}}=o\left(1\right). \tag{1.10.28}$$

Given equation (1.10.28), we apply the Bernstein inequality (e.g., Serfling, 1980, p.95, Masry 1996), then we have,

$$\sup_{v\in[-\gamma_0,\gamma_n]}p\left(\left|\sum_{i=1}^n w_{ni}\left(v\right)\right|>\eta\right)\leq 2\exp\left(-\lambda_n\eta+\frac{c_2\lambda_n^2}{nh}\right).$$

Let $\eta_n=c_3\left(\frac{\ln n}{nh}\right)^{\frac{1}{2}}$, $c_3>c_2$

$$\sup_{v\in[-\gamma_0,\gamma_n]}p\left(\left|\sum_{i=1}^n w_{ni}\left(v\right)\right|>\eta_n\right)\leq\frac{2}{n^{c_3-c_2}},$$

then combine equation (1.10.27), we can get

$$p\left(P_{12}>\eta_n\right)\leq\frac{2L_n}{n^{c_3-c_2}}. \tag{1.10.29}$$

By choosing $c_3 - c_2$ large enough, such that $\sum_{n=1}^{\infty} p\left(P_{12} > \eta_n\right) < \infty$. By Borel-Cantelli lemma, we know that

$$P_{12} = O_p\left(\left(\frac{\ln n}{nh}\right)^{\frac{1}{2}}\right). \tag{1.10.30}$$

By Lipschitz condition on $K\left(\cdot\right)$,

$$\sup_{v \in \mathbb{I}_{k,n}} \left| K\left(\frac{v - v_i}{h}\right) - K\left(\frac{v_{k,n} - v_i}{h}\right)\right| \leq \frac{c_K l_n}{h},$$

which implies that

$$|P_{11}| \leq \frac{c_4 l_n}{h^2}, \qquad |P_{13}| \leq \frac{c_5 l_n}{h^2},$$

for some $c_4, c_5 > 0$. Since by condition (1.3.5) and Lemma 1.10.1, $\gamma_n = O\left(n\right)$. Under the constraints for equation (1.10.29) and $l_n = \frac{2\gamma_n}{L_n}$, we could let $c_3 - c_2$ large enough and $L_n = n^{c_6}$ large enough such that $l_n = n^{-c_7}$ small enough such that

$$|P_{11}| = O_p\left(\left(\frac{\ln n}{nh}\right)^{\frac{1}{2}}\right), \qquad |P_{13}| = O_p\left(\left(\frac{\ln n}{nh}\right)^{\frac{1}{2}}\right). \tag{1.10.31}$$

From equation (1.10.30) and (1.10.31), we have

$$|P_1| \leq |P_{11}| + |P_{12}| + |P_{13}| = O_p\left(\left(\frac{\ln n}{nh}\right)^{\frac{1}{2}}\right). \tag{1.10.32}$$

By equation (1.10.26) and $q > \frac{1-c_h}{c_h}$, we have

$$P_2 = O_p\left(\left(\frac{\ln n}{nh}\right)^{\frac{1}{2}}\right). \tag{1.10.33}$$

The conclusion follows by equation (1.10.32) and (1.10.33). ■

Proof of Theorem 1.3.2.2 We can get that the residual term is $o_p\left(\sqrt{\frac{\sigma_n^2}{n}}\right)$ by Lemma 1.3.1 and results in Khan and Tamer (2009) Appendix B.2. Let

$$Q_n\left(z_i, z_j\right) = \frac{1}{2}\left(\frac{m_{ni}}{f_v^2\left(v_i\right)} + \frac{m_{nj}}{f_v^2\left(v_j\right)}\right)\frac{1}{h}K\left(\frac{v_j - v_i}{h}\right). \tag{1.10.34}$$

Then not hard to verify that

$$\begin{aligned}
&\frac{1}{n}\sum\nolimits_{i=1}^{n}\left(\frac{m_{ni}\widehat{f}_v\left(v_i\right)}{f_v^2\left(v_i\right)}\right) \\
&= \frac{1}{n\left(n-1\right)}\sum\nolimits_{i=1}^{n}\sum\nolimits_{j=1,j\neq i}^{n}\frac{m_{ni}}{f_v^2\left(v_i\right)}\frac{1}{h}K\left(\frac{v_j - v_i}{h}\right) \\
&= \frac{1}{n\left(n-1\right)}\sum\nolimits_{i=1}^{n}\sum\nolimits_{j=1,j\neq i}^{n}\frac{1}{2}\left(\frac{m_{ni}}{f_v^2\left(v_i\right)} + \frac{m_{nj}}{f_v^2\left(v_j\right)}\right)\frac{1}{h}K\left(\frac{v_j - v_i}{h}\right) \\
&= \frac{1}{n\left(n-1\right)}\sum\nolimits_{i=1}^{n}\sum\nolimits_{j=1,j\neq i}^{n}Q_n\left(z_i, z_j\right).
\end{aligned} \tag{1.10.35}$$

To apply U-statistics technique on equation (1.10.35), according to Powell et al. (1989), we need to verify $\mathbb{E}\left[Q_n^2\left(z_i, z_j\right)\right] = o\left(n\right)$. Since we know the rate of convergence here is $\sqrt{\frac{n}{\sigma_n^2}}$ instead of $\sqrt{n}$, not hard to see that we only need to verify a weaker condition $\mathbb{E}\left[Q_n^2\left(z_i, z_j\right)\right] = o\left(n\sigma_n^2\right)$.

First we need to get the order of $\mathbb{E}\left[Q_n^2\left(z_i, z_j\right)\right]$,

$$\mathbb{E}\left[Q_n^2\left(z_i, z_j\right)\right] \asymp \mathbb{E}\left[\frac{m_{ni}^2}{f_v^4\left(v_i\right)}\frac{1}{h^2}K^2\left(\frac{v_j - v_i}{h}\right)\right].$$

and

$$\mathbb{E}\left[\frac{m_{ni}^2}{f_v^4(v_i)}\frac{1}{h^2}K^2\left(\frac{v_j-v_i}{h}\right)\right]=\mathbb{E}\left[\frac{D_iT_{ni}Y_i^{*2}}{(\gamma_n-\mathbb{E}(U_n))^2 f_v^4(v_i)}\frac{1}{h^2}K^2\left(\frac{v_j-v_i}{h}\right)\right]$$
$$\asymp\int_{-\infty}^{\infty}\frac{p_D(v_i)T_{ni}}{\gamma_n^2 f_v^3(v_i)}\int_{-\infty}^{\infty}\frac{1}{h^2}K^2\left(\frac{v_j-v_i}{h}\right)f(v_j)\,dv_j dv_i$$
$$\asymp\int_{-1}^{1}K^2(u)\int_{-\gamma_0}^{\gamma_n}\frac{p_D(v_i)}{\gamma_n^2 f_v^2(v_i)}\frac{1}{h}\frac{f_v(v_i+hu)}{f_v(v_i)}dv_i du.$$

By Assumption 11 that $\frac{f_v(v+h)}{f_v(v)}=1+o(1)$, continue from last equality

$$\mathbb{E}\left[\frac{a_{1i}^2}{f_v^4(v_i)}\frac{1}{h^2}K^2\left(\frac{v_j-v_i}{h}\right)\right]\asymp\frac{1}{h\gamma_n^2}\int_{-\gamma_0}^{\gamma_n}\frac{p_D(v_i)}{f_v^2(v_i)}dv_i$$

Here we show

$$\frac{\mathbb{E}\left(Q_n^2(z_i,z_j)\right)}{n\sigma_n^2}=o(1),\tag{1.10.36}$$

where the left hand side is of the same order as

$$\frac{\gamma_n^2}{n\int_{-\gamma_0}^{\gamma_n}\frac{p_D(v)}{f_v(v)}dv}\frac{1}{h\gamma_n^2}\int_{-\gamma_0}^{\gamma_n}\frac{p_D(v)}{f_v^2(v)}dv=\frac{1}{nhf_v(\gamma_n)}\frac{\int_{-\gamma_0}^{\gamma_n}\frac{p_D(v)}{f_v(v)}\frac{f_v(\gamma_n)}{f_v(v)}dv}{\int_{-\gamma_0}^{\gamma_n}\frac{p_D(v)}{f_v(v)}dv}.$$

We already know that $nhf_v(\gamma_n)\to\infty$ and the other term on the right hand side is obviously bounded. Thus equation (1.10.36) holds.

Given equation (1.10.36), standard U-statistics result implies that[3]

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{m_{ni}\widehat{f}_v(v_i)}{f_v^2(v_i)}\right)=\mathbb{E}[Q_n(z_i,z_j)]+\frac{1}{n}\sum_{i=1}^{n}2\left(\mathbb{E}[Q_n(z_i,z_j)|z_i]-\mathbb{E}[Q_n(z_i,z_j)]\right)+o_p\left(\sqrt{\frac{\sigma_n^2}{n}}\right).\tag{1.10.37}$$

[3]It is got from taking expectation on the squared equation (1.10.35).

Note that

$$\begin{aligned}2\mathbb{E}\left[Q_n\left(z_i,z_j\right)\middle|z_i\right] &= \frac{m_{ni}}{f_v^2\left(v_i\right)}\mathbb{E}\left(\frac{1}{h}K\left(\frac{v_j-v_i}{h}\right)\middle|z_i\right)+\mathbb{E}\left(\frac{m_{nj}}{f_v^2\left(v_j\right)}\frac{1}{h}K\left(\frac{v_j-v_i}{h}\right)\middle|z_i\right)\\ &= \frac{m_{ni}}{f_v\left(v_i\right)}\int_{-1}^{1}K\left(u\right)\frac{f_v\left(v_i+hu\right)}{f_v\left(v_i\right)}du+\int_{-1}^{1}\frac{\mathbb{E}\left[m_{ni}\middle|v_i+hu\right]}{f_v\left(v_i+hu\right)}K\left(u\right)du\\ &= \frac{m_{ni}}{f_v\left(v_i\right)}+\frac{\mathbb{E}\left(m_{ni}\middle|v_i\right)}{f_v\left(v_i\right)}+R_{1i}=\Lambda_{ni}+\mathbb{E}\left(\Lambda_{ni}\middle|v_i\right)+R_{1i}\end{aligned}\tag{1.10.38}$$

where

$$R_{1i}=\Lambda_{ni}\left[\int_{-1}^{1}K\left(u\right)\frac{f_v\left(v_i+hu\right)}{f_v\left(v_i\right)}du-1\right]+\int_{-1}^{1}\left[\mathbb{E}\left(\Lambda_{ni}\middle|v_i+hu\right)-\mathbb{E}\left(\Lambda_{ni}\middle|v_i\right)\right]K\left(u\right)du.\tag{1.10.39}$$

And

$$\mathbb{E}\left[Q_n\left(z_i,z_j\right)\right]=\mathbb{E}\left(\Lambda_{ni}\right)+O\left(h^q\right).\tag{1.10.40}$$

by value of q and h specified in the lemma, easy to verify that $O\left(h^q\right)=O\left(n^{-\frac{1}{2}}\right)$.

Use equation (1.10.38)

$$2\left(\mathbb{E}\left[Q_n\left(z_i,z_j\right)\middle|z_i\right]-\mathbb{E}\left[Q_n\left(z_i,z_j\right)\right]\right)=\Lambda_{ni}+\mathbb{E}\left(\Lambda_{ni}\middle|v_i\right)-2\mathbb{E}\left(\Lambda_{ni}\right)+R_{1i}-\mathbb{E}\left(R_{1i}\right).\tag{1.10.41}$$

By Assumption 11,

$$R_{1i}=o_p\left(\Lambda_{ni}\right),\qquad R_{1i}^2=o_p\left(\Lambda_{ni}^2\right)$$

implying that

$$\frac{1}{n}\sum\nolimits_{i=1}^{n}\left(R_{1i}-\mathbb{E}\left(R_{1i}\right)\right)=o_p\left(\sqrt{\frac{\sigma_n^2}{n}}\right).\tag{1.10.42}$$

Combine equation (1.10.37), (1.10.40), (1.10.41), and (1.10.42), we have

$$\frac{1}{n}\sum\nolimits_{i=1}^{n}\left(\frac{m_{ni}\widehat{f}_v\left(v_i\right)}{f_v^2\left(v_i\right)}\right)=\frac{1}{n}\sum\nolimits_{i=1}^{n}\left[\Lambda_{ni}+\mathbb{E}\left(\Lambda_{ni}\middle|\, v_i\right)-\mathbb{E}\left(\Lambda_{ni}\right)\right]+o_p\left(\sqrt{\frac{\sigma_n^2}{n}}\right). \quad (1.10.43)$$

From equation (1.3.8) and that the residual term is asymptotic negligible, we have

$$\frac{1}{n}\sum\nolimits_{i=1}^{n}\widehat{\Lambda}_{ni}=\frac{1}{n}\sum\nolimits_{i=1}^{n}\left(2\Lambda_{ni}-\Lambda_{ni}-\mathbb{E}\left(\Lambda_{ni}\middle|\, v_i\right)+\mathbb{E}\left(\Lambda_{ni}\right)\right)+o_p\left(\sqrt{\frac{\sigma_n^2}{n}}\right).$$

Moving $\mathbb{E}\left(\Lambda_{ni}\right)$ from left hand side to the right hand side gives

$$\frac{1}{n}\sum\nolimits_{i=1}^{n}\left(\widehat{\Lambda}_{ni}-\mathbb{E}\left(\Lambda_{ni}\right)\right)=\frac{1}{n}\sum\nolimits_{i=1}^{n}\left(\Lambda_{ni}-\mathbb{E}\left(\Lambda_{ni}\middle|\, v_i\right)\right)+o_p\left(\sqrt{\frac{\sigma_n^2}{n}}\right),$$

which is the conclusion of the theorem. ■

Proof of Theorem 1.3.4.2 Not hard to see that

$$\mathbb{E}\left\{\left[\Lambda_{ni}-\mathbb{E}\left(\Lambda_{ni}|v_i\right)\right]^2\right\}\asymp\sigma_n^2,$$

so to show that one term is asymptotically negligible is equivalent to show that term is $o_p\left(\sqrt{\frac{\sigma_n^2}{n}}\right)$. From the expression (1.3.11), we know

$$\begin{aligned}
&\frac{1}{n}\sum\nolimits_{i=1}^{n}\left(\widehat{\Lambda}_{ni}^{*}-\frac{1}{n}\sum\nolimits_{i=1}^{n}\widehat{\Lambda}_{ni}\right)\\
&=\frac{1}{n}\sum\nolimits_{i=1}^{n}\left[\frac{2m_{ni}^{*}}{f_v\left(v_i^{*}\right)}-\frac{2m_{ni}}{f_v\left(v_i\right)}\right]-\frac{1}{n\left(n-1\right)}\sum\nolimits_{i=1}^{n}\sum\nolimits_{j=1,j\neq i}^{n}\left[Q_n\left(z_i^{*},z_j^{*}\right)-Q_n\left(z_i,z_j\right)\right]\\
&+\frac{1}{n}\sum\nolimits_{i=1}^{n}\left[\frac{m_{ni}^{*}\left(f_v\left(v_i^{*}\right)-\widehat{f}_v\left(v_i^{*}\right)\right)^2}{f_v^2\left(v_i^{*}\right)\widehat{f}_v\left(v_i^{*}\right)}-\frac{m_{ni}\left(f_v\left(v_i\right)-\widehat{f}_v\left(v_i\right)\right)^2}{f_v^2\left(v_i\right)\widehat{f}_v\left(v_i\right)}\right]. \quad (1.10.44)
\end{aligned}$$

Let

$$\Upsilon_1 = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{m_{ni}^{*}\left(f_v\left(v_i^{*}\right)-\widehat{f}_v\left(v_i^{*}\right)\right)^2}{f_v^2\left(v_i^{*}\right)\widehat{f}_v\left(v_i^{*}\right)}-\frac{m_{ni}\left(f_v\left(v_i\right)-\widehat{f}_v\left(v_i\right)\right)^2}{f_v^2\left(v_i\right)\widehat{f}_v\left(v_i\right)}\right],$$

then by Lemma 1.10.3, we show that $\Upsilon_1 = o_p\left(\sqrt{\frac{\sigma_n^2}{n}}\right)$.

For the U-statistics, Let

$$\Upsilon_2 = \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[Q_n\left(z_i^{*},z_j\right)-Q_n\left(z_i,z_j\right)-\mathbb{E}\left(Q_n\left(z_i^{*},z_j\right)\middle|z_i^{*}\right)+\mathbb{E}\left(Q_n\left(z_i,z_j\right)\middle|z_i\right)\right]$$

$$\Upsilon_3 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[Q_n\left(z_i^{*},z_j^{*}\right)-Q_n\left(z_i^{*},z_j\right)-Q_n\left(z_i,z_j^{*}\right)+Q_n\left(z_i,z_j\right)\right]$$

$$\Upsilon_4 = -\frac{1}{n^2}\sum_{i=1}^{n}\left[Q_n\left(z_i^{*},z_i^{*}\right)-Q_n\left(z_i,z_i\right)\right],$$

$$\Upsilon_5 = \frac{1}{n^2\left(n-1\right)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\left[Q_n\left(z_i^{*},z_j^{*}\right)-Q_n\left(z_i,z_j\right)\right],$$

then

$$\begin{aligned}
&\frac{1}{n\left(n-1\right)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\left[Q_n\left(z_i^{*},z_j^{*}\right)-Q_n\left(z_i,z_j\right)\right] \qquad (1.10.45)\\
&=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[Q_n\left(z_i^{*},z_j^{*}\right)-Q_n\left(z_i,z_j\right)\right]+\Upsilon_4+\Upsilon_5\\
&=\frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[Q_n\left(z_i^{*},z_j\right)-Q_n\left(z_i,z_j\right)\right]+\Upsilon_3+\Upsilon_4+\Upsilon_5.\\
&=\frac{2}{n}\sum_{i=1}^{n}\left[\mathbb{E}\left(Q_n\left(z_i^{*},z_j\right)\middle|z_i^{*}\right)-\mathbb{E}\left(Q_n\left(z_i,z_j\right)\middle|z_i\right)\right]+\Upsilon_2+\Upsilon_3+\Upsilon_4+\Upsilon_5,
\end{aligned}$$

where the third line holds because $Q_n\left(z_i,z_j\right)$ is symmetric in z_i,z_j. Υ_5 is obviously asymptotically negligible. Lemma 1.10.4 1.10.5 and 1.10.6 show that $\Upsilon_2 = o_p\left(\sqrt{\frac{\sigma_n^2}{n}}\right)$, $\Upsilon_3 = o_p\left(\sqrt{\frac{\sigma_n^2}{n}}\right)$, and $\Upsilon_4 = o_p\left(\sqrt{\frac{\sigma_n^2}{n}}\right)$, respectively.

Therefore, we have

$$\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\Lambda}_{ni}^{*}-\frac{1}{n}\sum_{i=1}^{n}\widehat{\Lambda}_{ni}\right)=\frac{1}{n}\sum_{i=1}^{n}\left[\frac{2m_{ni}^{*}}{f_v\left(v_i^{*}\right)}-\frac{2m_{ni}}{f_v\left(v_i\right)}\right. \tag{1.10.46}$$
$$\left.-2\mathbb{E}\left(Q_n\left(z_i^{*},z_j\right)\middle|\, z_i^{*}\right)+2\mathbb{E}\left(Q_n\left(z_i,z_j\right)\middle|\, z_i\right)\right]+o_p\left(\sqrt{\frac{\sigma_n^2}{n}}\right).$$

For the right hand side of equation (1.10.46), by equation (1.10.38), we know

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\left[\frac{2m_{ni}^{*}}{f_v\left(v_i^{*}\right)}-\frac{2m_{ni}}{f_v\left(v_i\right)}-2\mathbb{E}\left(Q_n\left(z_i^{*},z_j\right)\middle|\, z_i^{*}\right)+2\mathbb{E}\left(Q_n\left(z_i,z_j\right)\middle|\, z_i\right)\right]\\&=\frac{1}{n}\sum_{i=1}^{n}\left[\Lambda_{ni}^{*}-\mathbb{E}\left(\Lambda_{ni}^{*}\middle|\, v_i^{*}\right)-\left(\frac{1}{n}\sum_{i=1}^{n}\left[\Lambda_{ni}+\mathbb{E}\left(\Lambda_{ni}\middle|\, v_i\right)\right]\right)\right]+\frac{1}{n}\sum_{i=1}^{n}\left(R_{1i}^{*}-R_{1i}\right),\end{aligned}$$

where $\frac{1}{n}\sum_{i=1}^{n}\left(R_{1i}^{*}-R_{1i}\right)=o_p\left(\sqrt{\frac{\sigma_n^2}{n}}\right)$ for the same reason as equation (1.10.42). The Lindeberg condition for

$$\frac{1}{n}\sum_{i=1}^{n}\left[\Lambda_{ni}^{*}-\mathbb{E}\left(\Lambda_{ni}^{*}\middle|\, v_i^{*}\right)-\left(\frac{1}{n}\sum_{i=1}^{n}\left[\Lambda_{ni}+\mathbb{E}\left(\Lambda_{ni}\middle|\, v_i\right)\right]\right)\right]$$

hold by Assumptions in the theorem, and

$$\begin{aligned}&var\left(\frac{1}{n}\sum_{i=1}^{n}\left[\Lambda_{ni}^{*}-\mathbb{E}\left(\Lambda_{ni}^{*}\middle|\, v_i^{*}\right)-\left(\frac{1}{n}\sum_{i=1}^{n}\left[\Lambda_{ni}-\mathbb{E}\left(\Lambda_{ni}\middle|\, v_i\right)\right]\right)\right]\right)\\&=\frac{1}{n}\mathbb{E}\left\{\mathbb{E}\left[\left[\Lambda_{ni}^{*}-\mathbb{E}\left(\Lambda_{ni}^{*}\middle|\, v_i^{*}\right)\right]^2\middle|\, z_1,...,z_n\right]\right\}-\frac{1}{n}\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}\left[\Lambda_{ni}-\mathbb{E}\left(\Lambda_{ni}\middle|\, v_i\right)\right]\right)^2\right]\\&=\frac{1}{n}\mathbb{E}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[\Lambda_{ni}-\mathbb{E}\left(\Lambda_{ni}\middle|\, v_i\right)\right]^2\right\}-\frac{1}{n^2}\mathbb{E}\left\{\left[\Lambda_{ni}-\mathbb{E}\left(\Lambda_{ni}\middle|\, v_i\right)\right]^2\right\}\\&=\frac{1}{n}\mathbb{E}\left\{\left[\Lambda_{ni}-\mathbb{E}\left(\Lambda_{ni}\middle|\, v_i\right)\right]^2\right\}+o_p\left(\sqrt{\frac{\sigma_n^2}{n}}\right),\end{aligned} \tag{1.10.47}$$

where the first and second equalities hold because i.i.d. of $\{z_i^{*}\}_{i=1}^{n}$ conditional on $\{z_i\}_{i=1}^{n}$ and the i.i.d. of $\{z_i\}_{i=1}^{n}$ itself. Therefore, by equation (1.10.46) and (1.10.47), and Lindeberg-

Feller central limit theorem, we have

$$\sqrt{\frac{n}{\mathbb{E}\left\{\left[\Lambda_{ni}-\mathbb{E}\left(\Lambda_{ni}|v_i\right)\right]^2\right\}}}\left[\frac{1}{n}\sum\nolimits_{i=1}^{n}\left(\widehat{\Lambda}_{ni}^{*}-\frac{1}{n}\sum\nolimits_{i=1}^{n}\widehat{\Lambda}_{ni}\right)\right]\xrightarrow{d}N\left(0,1\right).$$

■

Lemma 1.10.2 *Under assumptions in Theorem 1.3.4,*

$$\frac{1}{n\sigma_n^2}\mathbb{E}\left[Q_n^2\left(z_i,z_j\right)\right]\to 0,\ \textit{for } i\neq j,\quad \frac{1}{n^2\sigma_n^2}\mathbb{E}\left[Q_n^2\left(z_i,z_i\right)\right]\to 0.$$

Proof of Lemma 1.10.2.2 The first conclusion is already shown in equation (1.10.36).

For the second conclusion, by some simple calculations,

$$\begin{aligned}\frac{1}{n^2\sigma_n^2}\mathbb{E}\left[Q_n^2\left(z_i,z_i\right)\right] &\asymp \frac{\gamma_n^2}{n^2\int_{-\gamma_0}^{\gamma_n}\frac{p_D(v)}{f_v(v)}dv}\frac{1}{\gamma_n^2h^2}\int_{-\gamma_0}^{\gamma_n}\frac{p_D\left(v\right)}{f_v^3\left(v\right)}dv\\ &=\frac{1}{n^2h^2f_v^2\left(\gamma_n\right)}\frac{\int_{-\gamma_0}^{\gamma_n}\frac{p_D(v)}{f_v(v)}\frac{f_v^2(\gamma_n)}{f_v^2(v)}dv}{\int_{-\gamma_0}^{\gamma_n}\frac{p_D(v)}{f_v(v)}dv},\end{aligned}$$

where $nhf_v\left(\gamma_n\right)\to\infty$ and $\frac{\int_{-\gamma_0}^{\gamma_n}\frac{p_D(v)}{f_v(v)}\frac{f_v^2(\gamma_n)}{f_v^2(v)}dv}{\int_{-\gamma_0}^{\gamma_n}\frac{p_D(v)}{f_v(v)}dv}$ is bounded. So we have $\frac{1}{n^2\sigma_n^2}\mathbb{E}\left[Q_n^2\left(z_i,z_i\right)\right]\to 0$.

■

Lemma 1.10.3 *Under assumptions in Theorem 1.3.4,* $\Upsilon_1=o_p\left(\sqrt{\frac{\sigma_n^2}{n}}\right)$.

Proof of Lemma 1.10.3.2 If we use $\{v_i^*\}_{i=1}^n$ to estimate f_v, then

$$\begin{aligned}\mathbb{E}\left[\widehat{f}_v(v)\right] &= \mathbb{E}\left[\frac{1}{nh}\sum\nolimits_{i=1}^n K\left(\frac{v_i^*-v}{h}\right)\right] = \mathbb{E}\left[\frac{1}{h}K\left(\frac{v_i^*-v}{h}\right)\right] \\ &= \mathbb{E}\left\{\mathbb{E}\left[\frac{1}{h}K\left(\frac{v_i^*-v}{h}\right)\middle| z_1,...,z_n\right]\right\} = \mathbb{E}\left\{\frac{1}{nh}\sum\nolimits_{i=1}^n K\left(\frac{v_i-v}{h}\right)\right\} \\ &= f_v(v) + \frac{\kappa_q}{p!}f_v^{(q)}(v)\,h^q.\end{aligned}$$

For the same reason,

$$var\left(\widehat{f}_v(v)\right) = \frac{\pi f_v(v)}{nh}.$$

Both terms coincide with equation (1.3.2) and (1.3.3).

Therefore, we could similarly prove that Lemma 1.3.1 hold for $\widehat{f}_v(v)$ using $\{v_i^*\}_{i=1}^n$. Apply the results in Khan and Tamer (2009) Appendix B.2., we have the conclusion. ■

Lemma 1.10.4 *Under assumptions in Theorem 1.3.4,* $\Upsilon_2 = o_p\left(\sqrt{\frac{\sigma_n^2}{n}}\right)$.

Proof of Lemma 1.10.4.2 Υ_2 could be rewritten as

$$\begin{aligned}\Upsilon_2 = \frac{2}{n}\sum\nolimits_{i=1}^n \Bigg[&\frac{1}{n}\sum\nolimits_{j=1}^n Q_n(z_i^*, z_j) - \mathbb{E}\left(Q_n(z_i^*, z_j)\middle| z_i^*\right) \\ &-\frac{1}{n^2}\sum\nolimits_{i=1}^n\sum\nolimits_{j=1}^n Q_n(z_i, z_j) + \frac{1}{n}\sum\nolimits_{i=1}^n \mathbb{E}\left(Q_n(z_i, z_j)\middle| z_i\right)\Bigg].\end{aligned}$$

Then not hard to check that $\mathbb{E}\left(\Upsilon_2\middle| z_1,...,z_n\right) = 0$, implying that $\mathbb{E}(\Upsilon_2) = 0$. For the second

moment of Υ_2, terms inside the first summation is i.i.d. given $z_1, ..., z_n$, so

$$\begin{aligned}\mathbb{E}\left(\Upsilon_2^2 \middle| z_1, ..., z_n\right) &= \frac{4}{n}\mathbb{E}\left\{\left[\frac{1}{n}\sum\nolimits_{j=1}^{n} Q_n\left(z_i^*, z_j\right) - \mathbb{E}\left(Q_n\left(z_i^*, z_j\right)\middle| z_i^*\right)\right]^2 \middle| z_1, ..., z_n\right\} \\ &\quad - \frac{4}{n}\left(\frac{1}{n^2}\sum\nolimits_{i=1}^{n}\sum\nolimits_{j=1}^{n}\left[Q_n\left(z_i, z_j\right) - \mathbb{E}\left(Q_n\left(z_i, z_j\right)\middle| z_i\right)\right]\right)^2 \\ &= \frac{4}{n}\left\{\frac{1}{n}\sum\nolimits_{i=1}^{n}\left[\frac{1}{n}\sum\nolimits_{j=1}^{n} Q_n\left(z_i, z_j\right) - \mathbb{E}\left(Q_n\left(z_i, z_j\right)\middle| z_i\right)\right]^2\right\} \\ &\quad - \frac{4}{n}\left(\frac{1}{n^2}\sum\nolimits_{i=1}^{n}\sum\nolimits_{j=1}^{n}\left[Q_n\left(z_i, z_j\right) - \mathbb{E}\left(Q_n\left(z_i, z_j\right)\middle| z_i\right)\right]\right)^2. \qquad (1.10.48)\end{aligned}$$

Easy to see that when $j \neq j'$,

$$\mathbb{E}\left\{\left[Q_n\left(z_i, z_j\right) - \mathbb{E}\left(Q_n\left(z_i, z_j\right)\middle| z_i\right)\right]\left[Q_n\left(z_i, z_{j'}\right) - \mathbb{E}\left(Q_n\left(z_i, z_{j'}\right)\middle| z_i\right)\right]\right\} = 0,$$

when $i \neq j \neq i' \neq j'$,

$$\mathbb{E}\left\{\left[Q_n\left(z_i, z_j\right) - \mathbb{E}\left(Q_n\left(z_i, z_j\right)\middle| z_i\right)\right]\left[Q_n\left(z_{i'}, z_{j'}\right) - \mathbb{E}\left(Q_n\left(z_{i'}, z_{j'}\right)\middle| z_{i'}\right)\right]\right\} = 0.$$

Therefore, taking unconditional expectation on equation (1.10.48), we can have

$$\mathbb{E}\left(\Upsilon_2^2\right) = \frac{c_1}{n^2}\mathbb{E}\left[Q_n^2\left(z_i, z_j\right)\right] + \frac{c_2}{n^3}\mathbb{E}\left[Q_n^2\left(z_i, z_i\right)\right], \qquad (1.10.49)$$

where $i \neq j$, and c_1, c_2 are some constants. By Lemma 1.10.2,

$$\mathbb{E}\left(\Upsilon_2^2\right) = o_p\left(\frac{\sigma_n^2}{n}\right)$$

implying the conclusion by Markov inequality. ■

Lemma 1.10.5 *Under assumptions in Theorem 1.3.4,* $\Upsilon_3 = o_p\left(\sqrt{\frac{\sigma_n^2}{n}}\right)$.

Proof of Lemma 1.10.5.2 We do this along the same line as in Lemma 1.10.4. First we rewrite Υ_3 as

$$\begin{aligned}\Upsilon_3 = \frac{1}{n^2}\sum_{i=1}^n \sum_{j=1}^n \Big[& Q_n\left(z_i^*, z_j^*\right) - \frac{1}{n}\sum_{j=1}^n Q_n\left(z_i^*, z_j\right) \\ & -\frac{1}{n}\sum_{j=1}^n Q_n\left(z_i, z_j^*\right) + \frac{1}{n^2}\sum_{i=1}^n \sum_{j=1}^n Q_n\left(z_i, z_j\right)\Big] \end{aligned} \tag{1.10.50}$$

Not hard to check that $\mathbb{E}\left(\Upsilon_3 \middle| z_1, ..., z_n\right) = 0$, implying that $\mathbb{E}\left(\Upsilon_3\right) = 0$. Note that

$$\mathbb{E}\left(\Upsilon_3^2 \middle| z_1, ..., z_n\right) = \frac{1}{n^4}\sum_{i=1}^n \sum_{j=1}^n \sum_{i'=1}^n \sum_{j'=1}^n \mathbb{E}\left(\Upsilon_{ij}^{(3)}\Upsilon_{i'j'}^{(3)} \middle| z_1, ..., z_n\right),$$

where $\Upsilon_{ij}^{(3)}$ is the i,j-th term inside equation (1.10.50). In the case when $(i,j) \neq (i',j')$ and $(i,j) \neq (j',i')$, easy to see that

$$\mathbb{E}\left(\Upsilon_{ij}^{(3)}\Upsilon_{i'j'}^{(3)} \middle| z_1, ..., z_n\right) = 0,$$

implying

$$\mathbb{E}\left(\Upsilon_{ij}^{(3)}\Upsilon_{i'j'}^{(3)}\right) = 0.$$

Therefore only n^2 terms left inside the quadra-summation, and then we have

$$\mathbb{E}\left(\Upsilon_3^2\right) = \frac{c_1}{n^2}\mathbb{E}\left[Q_n^2\left(z_i, z_j\right)\right] + \frac{c_2}{n^3}\mathbb{E}\left[Q_n^2\left(z_i, z_i\right)\right],$$

for some constants c_1, c_2, and $i \neq j$. By Lemma 1.10.2,

$$\mathbb{E}\left(\Upsilon_3^2\right) = o_p\left(\frac{\sigma_n^2}{n}\right)$$

implying the conclusion by Markov inequality. ■

Lemma 1.10.6 *Under assumptions in Theorem 1.3.4,* $\Upsilon_4 = o_p\left(\sqrt{\frac{\sigma_n^2}{n}}\right)$.

Proof of Lemma 1.10.6.2 Rewrite Υ_4 as

$$\Upsilon_4 = -\frac{1}{n^2}\sum\nolimits_{i=1}^{n}\left[Q_n\left(z_i^*, z_i^*\right) - \frac{1}{n}\sum\nolimits_{i=1}^{n} Q_n\left(z_i, z_i\right)\right].$$

Then for the same reason as in Lemma 1.10.4 and 1.10.5,

$$\mathbb{E}\left(\Upsilon_4^2\right) = \frac{c}{n^3}\mathbb{E}\left[Q_n^2\left(z_i, z_i\right)\right],$$

for some constant c. The conclusion follows similarly as in Lemma 1.10.4 and 1.10.5. ■

Proof of Theorem 1.4.1.2 Consider $\frac{1}{n-1}\sum_{j\neq i}\Lambda_{nj}^{(i)}\frac{1}{h}K\left(\frac{x_j - x_i}{h}\right)$, then the condition for denominator could be verified similarly. We take x_i as a constant here.

$$\begin{aligned}
&\mathbb{E}\left[\left(\Lambda_{nj}^{(i)}\frac{1}{h}K\left(\frac{x_j - x_i}{h}\right)\right)^2\middle| x_i\right] \\
&= \int_{\Omega_{x_j}}\int_{-\gamma_0}^{\gamma_n(x_i)} \frac{\mathbb{E}\left(D_j Y_j^{*2}\middle| x_j, v_j\right)}{h^2 f^2\left(v_j\middle| x_j\right)\left(\gamma_n\left(x_i\right) - \mathbb{E}\left(U_n|x_i\right)\right)^2} K^2\left(\frac{x_j - x_i}{h}\right) f\left(v_j\middle| x_j\right) dv_j f_x\left(x_j\right) dx_j\left(1 + o\left(1\right)\right) \\
&= \frac{\pi f_x\left(x_i\right)\left(1 + o\left(1\right)\right)}{h\gamma_n^2\left(x_i\right)}\int_{-\gamma_0}^{\gamma_n(x_i)} \frac{\mathbb{E}\left(D_i Y_i^{*2}\middle| x_i, v_i\right)}{f\left(v_i\middle| x_i\right)} dv_i \asymp \frac{1}{h\gamma_n^2\left(x_i\right)}\int_{-\gamma_0}^{\gamma_n(x_i)} \frac{p_{D|x_i}\left(v\right)}{f\left(v\middle| x_i\right)} dv.
\end{aligned}$$

To let Lindeberg condition hold, similar to Theorem 1.2.3, we analyze the following set

$$\begin{aligned}\Psi\left(\gamma_n\left(x_i\right)\right) &= \left\{ z_j \left| nf_v^2\left(v_j\middle| x_j\right) h \int_{-\gamma_0}^{\gamma_n(x_i)} \frac{\mathbb{E}\left(D_i Y_i^{*2}\middle| x_i\right)}{f\left(v_i\middle| x_i\right)} dv_i \right.\right. \\ &< \left.\left. \frac{cD_j T_{nj}^{(i)} Y_j^{*2}}{\varepsilon} K^2\left(\frac{x_j - x_i}{h}\right) \right\}\right. \\ &= \left\{ z_j \left| nf_v^2\left(v_j\middle| x_i\right)\left(1+o\left(1\right)\right) h \int_{-\gamma_0}^{\gamma_n(x_i)} \frac{\mathbb{E}\left(D_i Y_i^{*2}\middle| x_i\right)}{f\left(v_i\middle| x_i\right)} dv_i \right.\right. \\ &< \left.\left. \frac{cD_j T_{nj}^{(i)} Y_j^{*2}}{\varepsilon} K^2\left(\frac{x_j - x_i}{h}\right) \right\}\right., \end{aligned} \tag{1.10.51}$$

where c is some positive constant, and the second equality holds because the support of $K\left(.\right)$ is $[-1,1]$ and $f_v\left(v\middle| x+h\right) = f_v\left(v\middle| x\right)\left(1+o(1)\right)$. For the same reason as in Theorem 1.2.3, a sufficient condition is

$$nf_v^2\left(\gamma_n\left(x_i\right)\middle| x_i\right) h \int_{-\gamma_0}^{\gamma_n(x_i)} \frac{p_{D|x_i}\left(v\right)}{f\left(v\middle| x_i\right)} dv \to \infty. \tag{1.10.52}$$

The second conclusion in this Lemma follows similarly as in Theorem 1.2.9. ■

Lemma 1.10.7 *Suppose Assumption 35, 8, 12, 13, 14 15, and 37 hold. Let the residual term be*

$$\begin{aligned} R_{2i} = &- \frac{\left(\widehat{\mathbb{E}}\left(\Lambda_{ni}^{(i)}\middle| x_i\right) - \mathbb{E}\left(\Lambda_{ni}^{(i)}\middle| x_i\right)\right)\left(\widehat{\mathbb{E}}\left(\Pi_{ni}^{(i)}\middle| x_i\right) - \mathbb{E}\left(\Pi_{ni}^{(i)}\middle| x_i\right)\right)}{\mathbb{E}\left(\Pi_{ni}^{(i)}\middle| x_i\right)\widehat{\mathbb{E}}\left(\Pi_{ni}^{(i)}\middle| x_i\right)} \\ &+ \frac{\mathbb{E}\left(\Lambda_{ni}^{(i)}\middle| x_i\right)\left(\widehat{\mathbb{E}}\left(\Pi_{ni}^{(i)}\middle| x_i\right) - \mathbb{E}\left(\Pi_{ni}^{(i)}\middle| x_i\right)\right)^2}{\mathbb{E}\left(\Pi_{ni}^{(i)}\middle| x_i\right)^2 \widehat{\mathbb{E}}\left(\Pi_{ni}^{(i)}\middle| x_i\right)}. \end{aligned}$$

We choose bandwidth $h = n^{-c_h}$, for some $0 < c_h < c_h^*$, kernel function with order $q > \frac{1-c_h}{c_h}$

and γ_n from condition (1.4.7), for those with $f(v|x)$ that satisfy equation (1.4.6), we have

$$\sqrt{\frac{n}{\sigma_n^2}}\left(\frac{1}{n}\sum\nolimits_{i=1}^{n}\left|R_{2i}\right|\right)=o_p\left(1\right).$$

Proof of Lemma 1.10.7.2 By Assumption 14 that X lies in a compact set with a density bounded away from zero, modifying the results in Silverman (1978), and Li and Racine (2007) a little bit, we can have

$$\sup\left|\widehat{\mathbb{E}}\left(\Lambda_{ni}^{(i)}\middle|\, x_i\right)-\mathbb{E}\left(\Lambda_{ni}^{(i)}\middle|\, x_i\right)\right|=O\left(\sqrt{\frac{\ln\left(n\right)\widetilde{\sigma}_n^2\left(x_i\right)}{n}}\right),$$

$$\sup\left|\widehat{\mathbb{E}}\left(\Pi_{ni}^{(i)}\middle|\, x_i\right)-\mathbb{E}\left(\Pi_{ni}^{(i)}\middle|\, x_i\right)\right|=O\left(\sqrt{\frac{\ln\left(n\right)\widetilde{\sigma}_n^2\left(x_i\right)}{n}}\right).$$

Under condition (1.4.7), $\widehat{\mathbb{E}}\left(\Lambda_{ni}^{(i)}\middle|\, x_i\right)$ and $\widehat{\mathbb{E}}\left(\Pi_{ni}^{(i)}\middle|\, x_i\right)$ converge to $\mathbb{E}\left(\Lambda_{ni}^{(i)}\middle|\, x_i\right)$ and $\mathbb{E}\left(\Pi_{ni}^{(i)}\middle|\, x_i\right)$ which are $O_p\left(1\right)$. Therefore,

$$\begin{aligned}\sqrt{\frac{n}{\sigma_n^2}}\left(\frac{1}{n}\sum\nolimits_{i=1}^{n}\left|R_{2i}\right|\right)&=O_p\left(\sqrt{\frac{n}{\sigma_n^2}}\frac{\ln\left(n\right)\widetilde{\sigma}_n^2\left(x_i\right)}{n}\right)\\&=O_p\left(\left(\frac{\ln\left(n\right)^2\int_{-\gamma_0}^{\gamma_n(x_i)}\frac{p_{D|x_i}(v)}{f(v|x_i)}dv}{nh\gamma_n^2\left(x_i\right)}\right)^{\frac{1}{2}}\right)\\&=O_p\left(\left(\frac{\ln\left(n\right)^2}{nhf\left(\gamma_n\left(x_i\right)|\, x_i\right)}\frac{f\left(\gamma_n\left(x_i\right)|\, x_i\right)}{\gamma_n^2\left(x_i\right)}\int_{-\gamma_0}^{\gamma_n(x_i)}\frac{p_{D|x_i}\left(v\right)}{f\left(v|\, x_i\right)}dv\right)^{\frac{1}{2}}\right).\end{aligned}$$

Under condition (1.4.6), (1.4.7) and assumption on h, it is easy to verify that

$$\frac{\ln\left(n\right)^2}{nhf\left(\gamma_n\left(x_i\right)|\, x_i\right)}\to 0.$$

And the following is obvious

$$\frac{f\left(\gamma_n\left(x_i\right)|\, x_i\right)}{\gamma_n^2\left(x_i\right)} \int_{-\gamma_0}^{\gamma_n(x_i)} \frac{p_{\,D|x_i}\left(v\right)}{f\left(\,v|\, x_i\right)} dv \to 0.$$

So we have $\sqrt{\frac{n}{\sigma_n^2}}\left(\frac{1}{n}\sum_{i=1}^{n}\left|R_{2i}\right|\right) = o\left(1\right).$ ■

D Additional Tables and Pictures

Table 3: Symmetric Setting

Distribution of v		MEAN (TRUE=0)	STD	BIAS	RMSE
Panel A: $n = 200$, f_v known					
t(1)	Half Trim	-1.037	0.287	-1.037	1.076
	Full Trim	-0.409	0.376	-0.409	0.556
	Double Trim	-0.150	0.615	-0.150	0.633
	OLS	-0.755	0.200	-0.755	0.781
	Parametric	-0.001	0.330	-0.001	0.330
t(3)	Half Trim	-1.197	0.246	-1.197	1.222
	Full Trim	-0.561	0.376	-0.561	0.676
	Double Trim	-0.253	0.924	-0.253	0.958
	OLS	-0.910	0.191	-0.910	0.930
	Parametric	0.009	0.394	0.009	0.394
t(4)	Half Trim	-1.191	0.237	-1.191	1.214
	Full Trim	-0.582	0.384	-0.582	0.698
	Double Trim	-0.316	0.953	-0.316	1.004
	OLS	-0.934	0.192	-0.934	0.954
	Parametric	0.004	0.412	0.004	0.412
Panel B: $n = 1000$, f_v known					
t(1)	Half Trim	-0.667	0.140	-0.667	0.682
	Full Trim	-0.230	0.220	-0.230	0.318
	Double Trim	-0.075	0.373	-0.075	0.381
	OLS	-0.756	0.089	-0.756	0.761
	Parametric	0.000	0.145	0.000	0.145
t(3)	Half Trim	-0.984	0.110	-0.984	0.990
	Full Trim	-0.387	0.262	-0.387	0.467
	Double Trim	-0.156	0.828	-0.156	0.843
	OLS	-0.911	0.086	-0.911	0.915
	Parametric	-0.003	0.175	-0.003	0.175
t(4)	Half Trim	-1.013	0.105	-1.013	1.018
	Full Trim	-0.411	0.275	-0.411	0.494
	Double Trim	-0.202	0.892	-0.202	0.915
	OLS	-0.931	0.086	-0.931	0.935
	Parametric	-0.001	0.178	-0.001	0.178
Panel C: $n = 5000$, f_v known					
t(1)	Half Trim	-0.402	0.077	-0.402	0.409
	Full Trim	-0.137	0.127	-0.137	0.186
	Double Trim	-0.046	0.220	-0.046	0.225
	OLS	-0.755	0.040	-0.755	0.756
	Parametric	-0.001	0.065	-0.001	0.065

Table 3 (Continue): Symmetric Setting

Distribution of v		MEAN (TRUE=0)	STD	BIAS	RMSE
Panel C: $n = 5000$, f_v known (continue)					
t(3)	Half Trim	-0.770	0.057	-0.770	0.772
	Full Trim	-0.270	0.185	-0.270	0.327
	Double Trim	-0.115	0.708	-0.115	0.717
	OLS	-0.910	0.038	-0.910	0.911
	Parametric	0.000	0.077	0.000	0.077
t(4)	Half Trim	-0.825	0.053	-0.825	0.827
	Full Trim	-0.298	0.206	-0.298	0.362
	Double Trim	-0.139	0.847	-0.139	0.858
	OLS	-0.932	0.038	-0.932	0.933
	Parametric	-0.001	0.080	-0.001	0.080
Panel D: $n = 200$, f_v unknown					
t(3)	Half Trim	-1.058	0.221	-1.058	1.081
	Full Trim	-0.602	0.362	-0.602	0.702
	Double Trim	-0.333	0.856	-0.333	0.919
t(4)	Half Trim	-1.074	0.223	-1.074	1.097
	Full Trim	-0.617	0.369	-0.617	0.719
	Double Trim	-0.350	0.929	-0.350	0.992
Panel E: $n = 1000$, f_v unknown					
t(3)	Half Trim	-0.935	0.105	-0.935	0.941
	Full Trim	-0.408	0.263	-0.408	0.486
	Double Trim	-0.198	0.749	-0.198	0.774
t(4)	Half Trim	-0.973	0.103	-0.973	0.978
	Full Trim	-0.431	0.273	-0.431	0.510
	Double Trim	-0.230	0.874	-0.230	0.904
Panel F: $n = 5000$, f_v unknown					
t(3)	Half Trim	-0.771	0.057	-0.771	0.773
	Full Trim	-0.278	0.186	-0.278	0.334
	Double Trim	-0.140	0.654	-0.140	0.668
t(4)	Half Trim	-0.824	0.053	-0.824	0.826
	Full Trim	-0.305	0.205	-0.305	0.367
	Double Trim	-0.153	0.833	-0.153	0.847

Notes: True mean value is 0. MEAN, STD, BIAS, RMSE are the mean value, standard deviation, bias, and root mean square errors of the estimates, respectively.

Table 4: Asymmetric Setting

Distribution of v		MEAN (TRUE=0)	STD	BIAS	RMSE
Panel A: $n = 200$, f_v known					
t(1)	Half Trim	-0.757	0.365	-0.757	0.840
	Full Trim	-0.276	0.299	-0.276	0.407
	Double Trim	-0.063	0.348	-0.063	0.354
	OLS	-0.311	0.189	-0.311	0.363
	Parametric	-0.137	0.223	-0.137	0.261
t(3)	Half Trim	-0.837	0.426	-0.837	0.939
	Full Trim	-0.496	0.315	-0.496	0.587
	Double Trim	-0.141	0.608	-0.141	0.624
	OLS	-0.527	0.214	-0.527	0.569
	Parametric	-0.293	0.312	-0.293	0.428
t(4)	Half Trim	-0.836	0.417	-0.836	0.934
	Full Trim	-0.521	0.318	-0.521	0.610
	Double Trim	-0.174	0.628	-0.174	0.651
	OLS	-0.567	0.220	-0.567	0.608
	Parametric	-0.332	0.332	-0.332	0.469
Panel B: $n = 1000$, f_v known					
t(1)	Half Trim	-0.445	0.132	-0.445	0.464
	Full Trim	-0.083	0.148	-0.083	0.170
	Double Trim	-0.026	0.221	-0.026	0.223
	OLS	-0.311	0.082	-0.311	0.322
	Parametric	-0.136	0.097	-0.136	0.167
t(3)	Half Trim	-0.682	0.131	-0.682	0.694
	Full Trim	-0.183	0.209	-0.183	0.278
	Double Trim	-0.063	0.505	-0.063	0.509
	OLS	-0.527	0.096	-0.527	0.535
	Parametric	-0.288	0.135	-0.288	0.319
t(4)	Half Trim	-0.713	0.134	-0.713	0.726
	Full Trim	-0.219	0.234	-0.219	0.320
	Double Trim	-0.088	0.558	-0.088	0.565
	OLS	-0.566	0.098	-0.566	0.575
	Parametric	-0.330	0.147	-0.330	0.361
Panel C: $n = 5000$, f_v known					
t(1)	Half Trim	-0.186	0.060	-0.186	0.195
	Full Trim	-0.045	0.078	-0.045	0.090
	Double Trim	-0.015	0.122	-0.015	0.123
	OLS	-0.313	0.037	-0.313	0.315
	Parametric	-0.137	0.043	-0.137	0.143

Table 4 (Continue): Asymmetric Setting

Distribution of v		MEAN (TRUE=0)	STD	BIAS	RMSE
Panel C: $n = 5000$, f_v known (continue)					
t(3)	Half Trim	-0.551	0.060	-0.551	0.555
	Full Trim	-0.111	0.131	-0.111	0.171
	Double Trim	-0.031	0.413	-0.031	0.414
	OLS	-0.526	0.043	-0.526	0.528
	Parametric	-0.288	0.061	-0.288	0.294
t(4)	Half Trim	-0.605	0.059	-0.605	0.608
	Full Trim	-0.132	0.154	-0.132	0.203
	Double Trim	-0.051	0.501	-0.051	0.504
	OLS	-0.567	0.044	-0.567	0.569
	Parametric	-0.329	0.064	-0.329	0.335
Panel D: $n = 200$, f_v unknown					
t(3)	Half Trim	-0.838	0.430	-0.838	0.942
	Full Trim	-0.488	0.319	-0.488	0.583
	Double Trim	-0.175	0.550	-0.175	0.577
t(4)	Half Trim	-0.834	0.428	-0.834	0.938
	Full Trim	-0.514	0.318	-0.514	0.604
	Double Trim	-0.214	0.583	-0.214	0.621
Panel E: $n = 1000$, f_v unknown					
t(3)	Half Trim	-0.671	0.127	-0.671	0.682
	Full Trim	-0.185	0.212	-0.185	0.282
	Double Trim	-0.083	0.463	-0.083	0.471
t(4)	Half Trim	-0.707	0.132	-0.707	0.719
	Full Trim	-0.226	0.239	-0.226	0.329
	Double Trim	-0.107	0.547	-0.107	0.558
Panel F $n = 5000$, f_v unknown					
t(3)	Half Trim	-0.545	0.061	-0.545	0.549
	Full Trim	-0.111	0.133	-0.111	0.174
	Double Trim	-0.044	0.402	-0.044	0.404
t(4)	Half Trim	-0.601	0.060	-0.601	0.604
	Full Trim	-0.134	0.155	-0.134	0.205
	Double Trim	-0.064	0.488	-0.064	0.492

Notes: True mean value is 0. MEAN, STD, BIAS RMSE are the mean value, standard deviation, bias, and root mean square errors of the estimates, respectively.

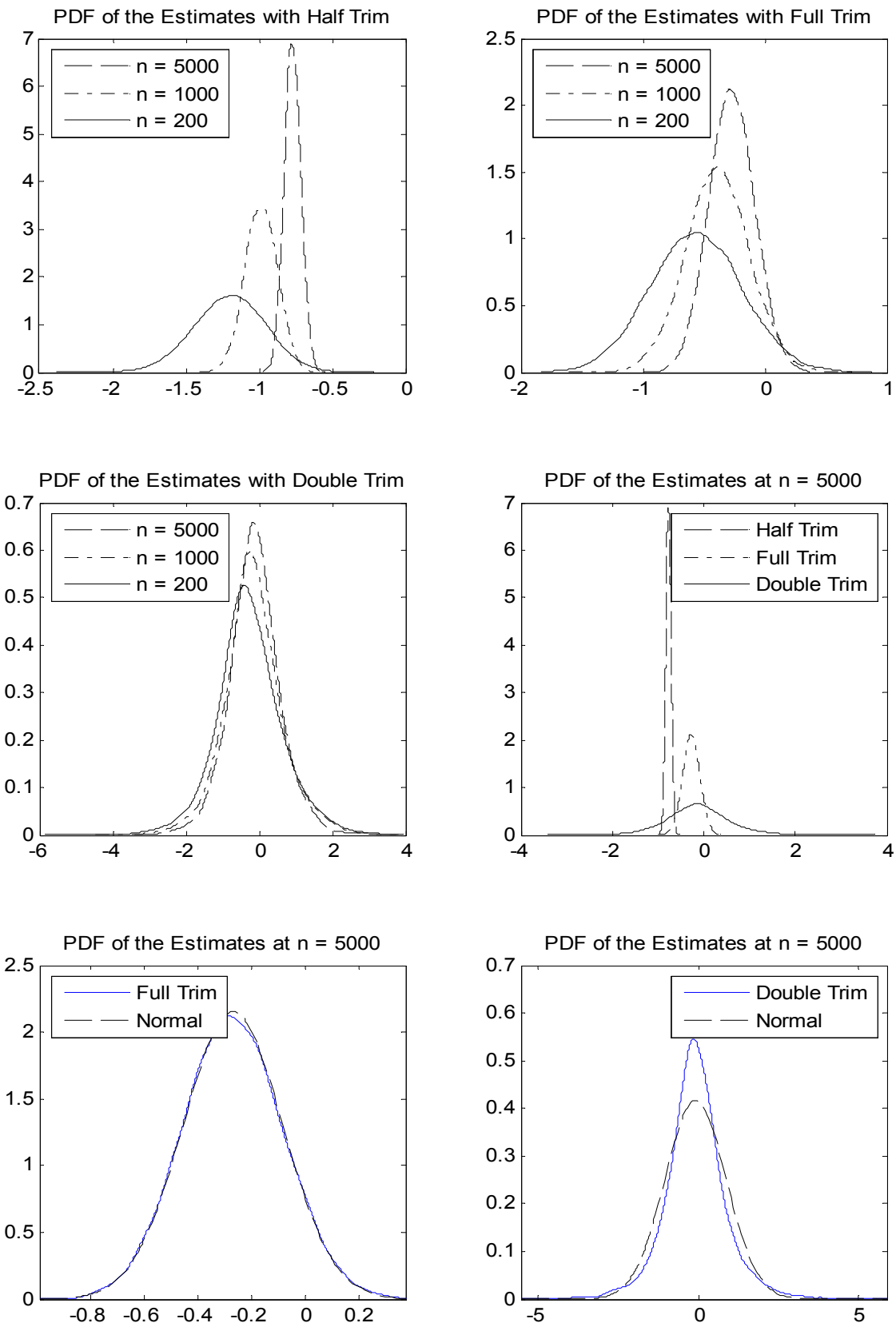

Figure 1.2: Monte Carlo Results in the First Experiment with V Distributed as t(3)

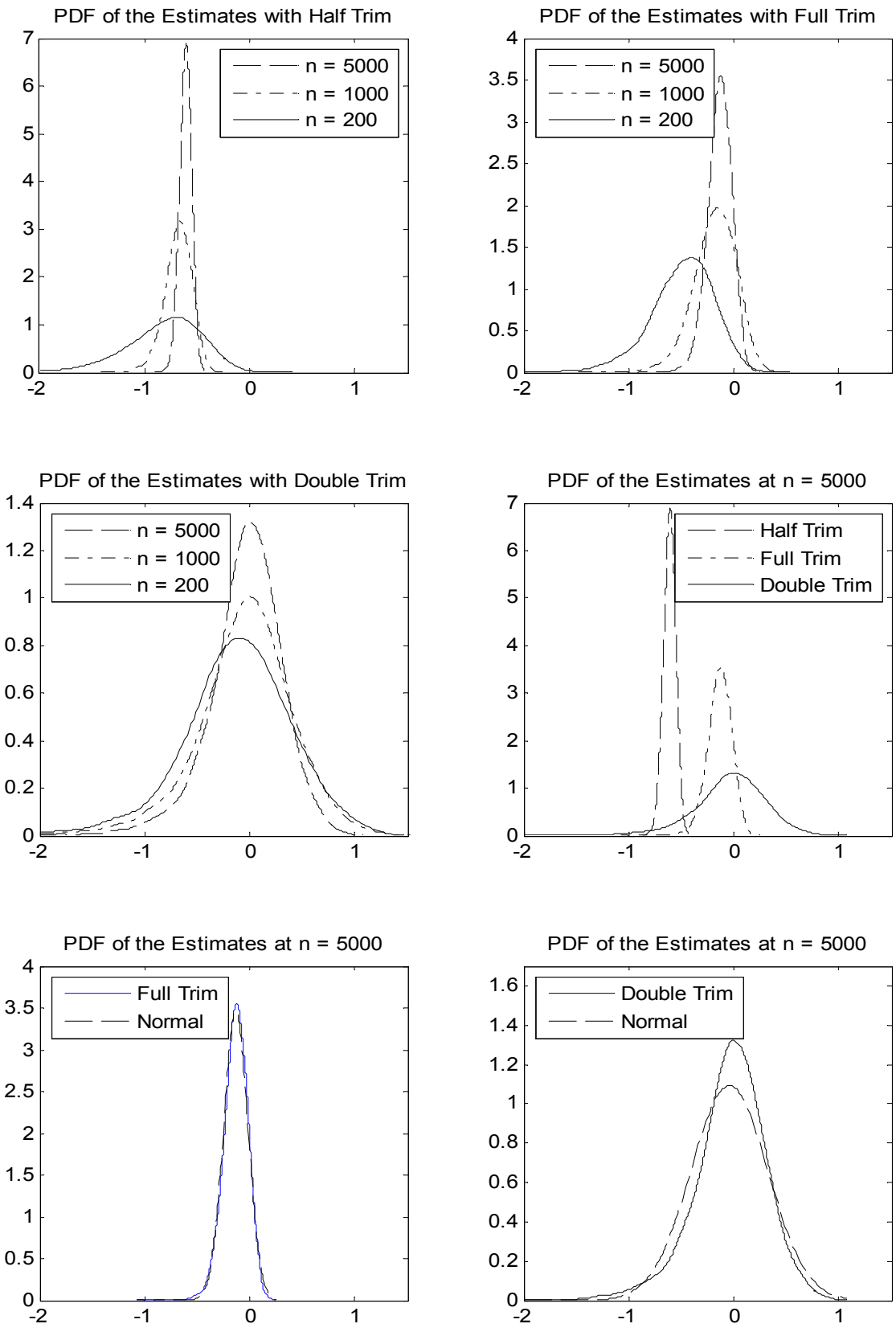

Figure 1.3: Monte Carlo Results in the Second Experiment with V Distributed as t(3)

Chapter 2

Identifying the Average Treatment Effect in a Two Threshold Model

With Arthur Lewbel

2.1 Introduction

Suppose an outcome Y is given by

$$Y = Y_0 + (Y_1 - Y_0) D \tag{2.1.1}$$

where Y_0 and Y_1 are potential outcomes as in Rubin (1974), and D is a binary treatment indicator. Generally, point identification of the average treatment effect (ATE) $E(Y_1 - Y_0)$ requires either i) conditional or unconditional unconfoundedness, or ii) an instrument for D that can drive D to zero and to one (with probability one), or iii) functional restrictions on the joint distribution of Y_0, Y_1 and D. In contrast, we provide a novel point identification

result, and an associated estimator, for the ATE in a model where none of these conditions hold.

Let V be a continuous instrument that affects the probability of treatment but not the outcome, and let X denote a vector of other covariates. In our model, D is given by a structure that is identical to one of the middle choices in an ordered choice model, that is,

$$D = I\left[\alpha_0(X) \leq V + U \leq \alpha_1(X)\right] \tag{2.1.2}$$

where $I(\cdot)$ is the indicator function that equals one if $\cdot$ is true and zero otherwise, U is a latent error term, and $\alpha_0(X)$ and $\alpha_1(X)$ are unknown functions. The joint distribution of $(U, Y_0, Y_1 \mid X)$ is assumed to be unknown.

In the special case of this model where $\alpha_0(X)$ and $\alpha_1(X)$ are linear with the same slope, this is equivalent to treatment being given by the more standard looking ordered choice specification

$$D = I\left(\delta_0 \leq X'\beta_1 + V + U \leq \delta_1\right)$$

for constants δ_0, δ_1, and β_1. However, we don't impose these linearity restrictions. In addition, unlike standard ordered choice models, we allow the distribution of U to depend on X in completely unknown ways. Equivalently, the covariates X can all be endogenous regressors, with no available associated instruments. The only covariate we require to be exogenous is V.

The proposed model is confounded, because the unobservable U that affects D can be correlated with Y_0 and Y_1, with or without conditioning on X. No parametric or semi-parametric restrictions are placed on the distribution of $(U, Y_0, Y_1 \mid X)$, so treatment effects

are not identified by functional form restrictions on the distributions of unobservables. We assume V has large support, but the model is not identified at infinity. This is because both very large and very small values of V drive the probability of treatment close to zero, but no value of V (or of other covariates) drives the probability of treatment close to one. So in this framework none of the conditions that are known to permit point identification of the ATE hold. Even a local ATE (LATE) is not identified in the usual way, because monotonicity of treatment with respect to the instrument cannot hold in the proposed model. Nevertheless, we show that the ATE is identified in our model, using a special regressor argument as in Lewbel (1998, 2000a, 2007). We also provide conditions under which a corresponding simple estimator of the ATE converges at rate root n.

To illustrate the model and foreshadow our later empirical application, suppose the outcome Y is a measure of innovation in an industry and $D = 1$ when a latent measure of competitiveness in the industry lies between two estimated thresholds, otherwise $D = 0$. According to the "Inverted-U" theory in Aghion, Bloom, Blundell, Griffith, and Howitt (2005) (hereafter ABBGH), industries with intermediate levels of competitiveness have more innovation than those with low levels or high levels of competition. As in Revenga (1990, 1992), Bertrand (2004), and Hashmi (2013), we use a source-weighted average of industry exchange rates as an instrumental variable for competition, which we take to be our special regressor V. This instrument is computed from the weighted average of the US dollar exchange rate with the currencies of its trading partners. When V is low, products from the U.S. becomes relatively cheaper, thereby reducing competition by driving out competitors. The treatment effect we estimate is therefore the gains in innovation that result from facing moderate (rather than low or high) levels of competition.

More generally, our estimator is potentially useful in applications where one wants to assess the impact of a treatment defined as a moderate level of some activity, versus low or high levels. Many such treatments exist. For example, one might want to assess the effects of moderate levels of BMI or of alcohol consumption on a variety of health outcomes (see, e.g., Cao et al. 2014, Koppes et al. 2005, and Solomon et al. 2000). Other examples are the effect of moderate levels of financial development on the growth rates of countries (see Cecchetti and Kharroubi 2012) or the effects of moderate levels of financial regulation on measures of financial instability (see Huang 2015).

Often one might be interested in comparing mean outcomes of the middle group, Y_1, with just the high group (those above the upper threshold) or just the low group (those below the lower threshold). We provide an extension of our results that combines our main identification theorem with identification at infinity arguments as in Heckman, Urzua and Vytlacil (2006) to estimate these additional treatment effects. This would be useful for applications such as returns to education, where, e.g., treatment could correspond to not finishing high school (the low group), finishing high school (the middle group), and having some college (the high group). Another extension we consider is identification in a model where V in the treatment equation is replaced with $\varsigma(V)$ for some unknown function ς.

Our empirical application uses panel data. We extend our method to show identification of $E(Y_{1it} - Y_{0it})$ in the panel data model

$$Y_{it} = \widetilde{a}_i + \widetilde{b}_t + Y_{0it} + (Y_{1it} - Y_{0it})D_{it}, \tag{2.1.3}$$

$$D_{it} = I(\alpha_0(x_{it}) \leq a_i + b_t + V_{it} + U_{it} \leq \alpha_1(x_{it})), \tag{2.1.4}$$

where $a_i, \widetilde{a}_i, b_t, \widetilde{b}_t$ are individual and time dummies in selection and outcome equations. If potential outcomes are given by $\widetilde{Y}_{dit} = \widetilde{a}_i + \widetilde{b}_t + Y_{dit}$, then our estimand $E\left(Y_{1it} - Y_{0it}\right)$ equals a standard ATE $E\left(\widetilde{Y}_{1it} - \widetilde{Y}_{0it}\right)$. Alternatively, given equation (2.1.3), $E\left(Y_{1it} - Y_{0it}\right)$ can be interpreted as a generalization of difference-in-difference (DID) estimation, where unlike standard DID, here D_{it} can be endogenous and hence correlated with the potential outcomes, so unconfoundedness does not hold. Equation (2.1.3) is also a generalization of Manski and Pepper (2003).[1] Despite the presence of fixed effects (incidental parameters) in the nonlinear selection equation, we attain a rate root nT estimate for the ATE in this panel model. We also consider other panel specifications, including dynamic panels.

The next section is a literature review. In section 3 we provide formal assumptions of our model, prove identification, and establish the consistency and asymptotic normality of our cross section and panel estimators. In section 4 we empirically apply our estimator to investigate the relationship between competition and innovation. In this section we also implement simulation experiments to evaluate small sample properties of our estimators, using a Monte Carlo design that replicates features of our empirical data. This is followed by an extensions section and conclusions. The paper additionally includes some appendices. Appendix A provides an evaluation of how the robustness of our approach compares to more structural models in the presence of measurement errors. Appendix B provides some additional extensions, and Appendix C gives additional technical assumptions and proofs. Finally, in a supplemental appendix separate from the main paper, we provide more details regarding application of relatively standard semiparametric methods for deriving the

[1] Manski and Pepper (2003) consider the linear treatment response model $Y_{it} = \alpha_i + \beta D_{it} + \gamma t + \varepsilon_{it}$ where α_i is the individual fixed effect, γt is the time trend, ε_{it} is the random disturbance, and β defines the ATE. Our model generalizes theirs by replacing their fixed ATE β with a random coefficient and replacing the time trend γt with time fixed effects $\widetilde{b}_t$.

limiting distribution of our estimators, and other technical material.

2.2 Literature Review

Existing methods for point identifying ATE's are discussed in surveys such as Heckman and Vytlacil (2007a, 2007b) and Imbens and Wooldridge (2009). The early treatment effects literature achieves identification by assuming unconfoundedness, see, e.g., Cochran and Rubin (1973), Rubin (1974), Barnow, Cain, and Goldberger (1980), Rosenbaum and Rubin (1983), and Heckman and Robb (1984). As noted by ABBGH, competition is an endogenous regressor, e.g., successful innovations increase market power and may thereby reduce competition. Much of what determines both is difficult to observe or even define, making it very unlikely that unconfoundedness would hold, regardless of what observable covariates one conditions upon.

Without unconfoundedness, instrumental variables have been used in a variety of ways to identify treatment effects. Instead of estimating the ATE, Imbens and Angrist (1994) show identification of a local average treatment effect (LATE), which is the ATE for a subpopulation called compliers (the definition of who compliers are, and hence the LATE, depends on the choice of instrument). An assumption for identifying the LATE is that the probability of treatment increase monotonically with the instrument. This assumption does not hold in our application, since both increasing or decreasing V sufficiently causes the probability of treatment to decrease. Although he does not provide an example, an implication of Kitagawa (2009) is that, if point identification of the ATE based only on an exogenous instrument were possible without identification at infinity, then instrument nonmonotonicity would be necessary. Our model possesses this necessary nonmonotonicity.

Building on Björklund and Moffitt (1987), Heckman and Vytlacil (1999, 2005, 2007a) describe identification of a marginal treatment effect (MTE) as a basis for program evaluation. The MTE is based on having a continuous instrument, as we do. However, identification of the ATE using the MTE requires the assumption that variation in V can drive the probability of treatment to either zero or one, and hence depends on an identification at infinity argument. As we have already noted, identification at infinity is not possible in our model, since no value of V can drive the probability of treatment to one.

A few other papers consider identification of treatment effects in ordered choice models, such as Angrist and Imbens (1995) and Heckman, Urzua, and Vytlacil (2006). However, these papers deal with models having more information than ours, i.e, observing extreme as well as middle choices, and they consider identification of LATE and MTE, respectively, not ATE. In an extension section, we will consider combining the information obtained by these approaches with our estimator.

The way we achieve identification here is based on special regressor methods, particularly Lewbel (2007), which exploits a related result to identify a class of semiparametric selection models. The instrumental variable V needs to be continuous, conditionally independent of other variables and have a large support, which are all standard assumptions for special regressor based estimators. See, e.g., Dong and Lewbel (2015), Lewbel, Dong, and Yang (2012), and Lewbel (2012). Some of the previously discussed papers also implicitly assume a special regressor, notably, Heckman, Urzua, and Vytlacil (2006).

In addition to the ATE, our methods can be immediately extended to estimate quantile treatment effects as in Abadie, Angrist, and Imbens (2002), Chernozhukov and Hansen (2005). Bitler, Gelbach, and Hoynes (2006), or Firpo (2006). This is done by replacing Y

with $I(Y \leq y)$ in our estimator.

In the panel context of equations (2.1.3) and (2.1.4), if unconfoundedness held so that $(Y_{0it}, Y_{1it}) \perp D_{it} \mid X_{it}$, and if in addition a_i and b_t were absent from the selection equation, then one could achieve identification via difference-in-difference methods, as in Ashenfelter (1978), Ashenfelter and Card (1985), Cook and Tauchen (1982, 1984), Card (1990), Meyer, Viscusi, and Durbin (1995), Card and Krueger (1993, 1994) and many others. In contrast, we obtain identification without unconfoundedness, and while allowing for a_i and b_t fixed effects. Analogous to Honore and Lewbel (2002), in panel data our identification and estimation strategy overcomes the incidental parameters problem associated with these fixed effects, and we attain a rate root nT estimate for the ATE.

Chernozhukov et al. (2009) discuss partial identification of marginal effects in nonlinear panel data, while Manski and Pepper (2013) provide partial identification of the (ATE) in a panel data context. Manski and Pepper also consider additional assumptions needed for point identification of the ATE in a panel setting (see their section 3.1). Our panel data point identification requires some but not all of the assumptions they list as needed, including an average treatment response that is time-invariant, and the instrument exclusion restriction in the outcome equation.

2.3 The Model

In this section we first prove identification of the ATE in our model. The proof we provide is constructive, and we next describe a corresponding estimator. This is followed by some extensions, in particular, a panel data estimator with fixed effects. The remaining parts of this section then provide limiting distribution theory for the estimators.

2.3.1 Identification and Estimation

Let $\Omega_\cdot$ and $f_\cdot$ denote supports and density functions for random variable $\cdot$, e.g., Ω_x and f_x are the support and density function for the random variable X. Let $\widehat{E}(\cdot)$ denote the sample mean of the argument inside, and let $\widehat{f}(\cdot)$ and $\widehat{E}(\cdot|\cdot)$ denote nonparametric Nadayara-Watson kernel density and kernel regression estimators, with bandwidth denoted h. For notational convenience, h is assumed the same for all covariates. We use R to denote any set of residual terms that are proven to be asymptotically negligible for our derived limiting distributions.

Assumption 18 *We observe realizations of an outcome Y, binary treatment indicator D, a covariate V, and a $k \times 1$ covariate vector X. Assume the outcome Y and treatment indicator D are given by equations (2.1.1) and (2.1.2) respectively, where $\alpha_0(X)$ and $\alpha_1(X)$ are unknown threshold functions with $\alpha_0(X) < \alpha_1(X)$, U is an unobserved latent random error, and Y_0 and Y_1 are unobserved random untreated and treated potential outcomes. The joint distribution of (U, Y_0, Y_1), either unconditional or conditional on X, is unknown.*

Assumption 19 *Assume $E(Y_j|X,V,U) = E(Y_j|X,U)$ for $j = 0, 1$, and $V \perp U \mid X$. Assume $V \mid X$ is continuously distributed with probability density function $f(V \mid X)$. For all $x \in supp(X)$, the $supp(V \mid X = x)$ is an interval on the real line, and the interval $[\inf supp(\alpha_0(X) - U \mid X = x), \sup supp(\alpha_1(X) - U \mid X = x)]$ is contained in $supp(V \mid X = x)$.*

Assumption 18 defines the model, while Assumption 19 says that V is an instrument, in that V affects the probability of treatment but not outcomes (after conditioning on X). The instrument V is also continuously distributed, and has a large enough support so that, for any values U and X may take on, V can be small enough to make $D = 0$ or large enough

to make $D = 0$. But no value of V and X will force $D = 1$, so identification at infinity is not possible.[2]

Remark 2.3.1 For identification, the assumption that $supp(V \mid X = x)$ equals an interval can be relaxed, as long as this support suitably contains $\alpha_0(x) - U$ and $\alpha_1(x) - U$ for all x. We maintain the single interval support to simplify notation in the identification proofs, and to apply the testing results in Section 2.5.1.

In this model, obtaining identification by imposing unconfoundedness would be equivalent to assuming that U was independent of $Y_1 - Y_0$, possibly after conditioning on covariates X. However, we do not make any assumption like this, so unconfoundedness does not hold. Alternatively, one might parametrically model the dependence of $Y_1 - Y_0$ on U to identify the model. In contrast we place no restrictions on the joint distribution of (U, Y_0, Y_1), either unconditional or conditioning upon X.

Assumption 20 *For some positive constant τ, define the trimming function $I_\tau(v, x) = I[\inf supp(V|X = x)+\tau \leq v \leq \sup supp(V|X = x)-\tau]$. Assume the interval $[\inf supp(\alpha_0(X) - U \mid X = x), \sup supp(\alpha_1(X) - U \mid X = x)]$ is contained in $\{v : I_\tau(v, x) = 1\}$.*

Assumption 21 *Assume there exists a positive constant $\widetilde{\tau} < \tau$ such that, for all v,x having $I_{\widetilde{\tau}}(v, x) = 1$, the density $f(v|x)$ is bounded away from zero (except possibly on a set of measure zero) and is bounded.*

Assumption 20 is not necessary for identification, but will be convenient for simplifying the limiting distribution theory for the estimator we construct based on the identification. In

[2] If instead of the ordered choice $D = I[\alpha_0(X) \leq V + U \leq \alpha_1(X)]$ we had a threshold crossing binary choice $D = I(\alpha_0(X) \leq V + U)$, then Assumption 19 would suffice to use "identification at infinity" to identify the treatment effect, by using data where V was arbitrarily low to estimate $E(Y_0 \mid X)$ and data where V was arbitrarily high to estimate $E(Y_1 \mid X)$. However, in our ordered choice model identification at infinity is not possible, since no value of V guarantees with high probability that Y will equal Y_1.

particular, this assumption permits fixed trimming that avoids boundary bias in our kernel estimators. This assumption could be relaxed using asymptotic trimming arguments. The requirement that $f(v|x)$ is bounded away from zero in Assumption 21 might also be relaxed via asymptotic trimming (e.g., by including another trimming indicator $I\left(f(v|x) > b_n\right)$, $b_n \to 0$, as $n \to \infty$). To save notation, we let $I_\tau \equiv I_\tau(V, X)$. Define the function $\psi\left(X\right)$ by

$$\psi\left(X\right) \equiv \frac{E\left[I_\tau DY/f(V \mid X) \mid X\right]}{E\left[I_\tau D/f(V \mid X) \mid X\right]} - \frac{E\left[I_\tau \left(1-D\right) Y/f(V \mid X) \mid X\right]}{E\left[I_\tau \left(1-D\right)/f(V \mid X) \mid X\right]} \tag{2.3.1}$$

Theorem 2.3.2 *Let Assumptions 18, 19 hold with $I_\tau = 1$, or let Assumptions 18, 19 20 and 21 hold. Then*

$$\psi\left(X\right) = E\left(Y_1 - Y_0 \mid X\right)$$

The theorem is proved in Appendix C. Theorem 2.3.2 is related to Lewbel (2007), however, that paper estimates a semiparametric selection model, while we identify and estimate a nonparametric conditional treatment effect. This includes identification for the untreated $E\left(Y_0|X\right)$ which is not considered in Lewbel (2007). We later provide more results that do not have analogs in Lewbel (2007), including, in Section 2.3.3, identification of a panel data model with fixed effects.

Theorem 2.3.2 shows identification of the conditional ATE since $\psi\left(X\right)$ is defined in terms of moments and densities of observed variables. The first part of the Theorem shows that just Assumptions 18 and 19 are needed for identification. The second part of the Theorem, giving identification including the additional Assumptions 20 and 21, is convenient because inclusion of the trimming term I_τ simplifies the asymptotics of the associated estimator.

It follows immediately from Theorem 2.3.2 that $\Psi \equiv E\left[\psi\left(X\right)\right]$ equals the ATE, which is therefore identified and can be consistently estimated by $\widehat{\Psi} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\psi}\left(x_i\right)$ where

$$\widehat{\psi}\left(x\right) = \frac{\widehat{E}\left[I_\tau DY/\widehat{f}(V \mid X) \mid X = x\right]}{\widehat{E}\left[I_\tau D/\widehat{f}(V \mid X) \mid X = x\right]} - \frac{\widehat{E}\left[I_\tau\left(1-D\right)Y/\widehat{f}(V \mid X) \mid X = x\right]}{\widehat{E}\left[I_\tau\left(1-D\right)/\widehat{f}(V \mid X) \mid X = x\right]},$$

with uniformly consistent kernel estimators $\widehat{f}$ and $\widehat{E}$.

To provide some intuition for Theorem 2.3.2, suppose for the moment that X was empty, and consider

$$E\left(D \mid U, Y_0, Y_1\right) = E\left(I\left[\alpha_0 - U \leq V \leq \alpha_1 - U\right] \mid U, Y_0, Y_1\right)$$

$$= \int_{supp(V)} I\left[\alpha_0 - U \leq v \leq \alpha_1 - U\right] f\left(v \mid U, Y_0, Y_1\right) dv = \int_{\alpha_0 - U}^{\alpha_1 - U} f\left(v|Y_0, Y_1\right) dv = F_{v|Y_0,Y_1}\left(\alpha_1 - U\right) - F_{v|Y_0,Y_1}\left(\alpha_0\right.$$

where $F_{v|Y_0,Y_1}$ is the cumulative density function of V conditional on Y_0, Y_1. We have confoundedness because the above expression depends on U, which is correlated with Y_0 and Y_1. However, if V were uniformly distributed, then the above expression would simplify to $E\left(D \mid U, Y_0, Y_1\right) = \alpha_1 - \alpha_0$, which is independent of (U, Y_0, Y_1). So if V were uniformly distributed, the model would be unconfounded. Moreover, in that case f would be constant and equation (2.3.1) would reduce to the standard propensity score weighted estimator of the (unconfounded) average treatment effect. Scaling by the density of V in equation (2.3.1) is equivalent to converting to a uniform V, and so is equivalent to converting our model into one that is unconfounded. Density weighting is a feature of some special regressor estimators including Lewbel (2000a, 2007), and indeed V has the properties of a special regressor, including appearing additively to unobservables in the model, a continuous distribution,

large support, and conditional independence.

2.3.2 Small Extensions

The above identification and associated estimator can be extended to handle independent random thresholds, that is, all the results go through if the deterministic functions $\alpha_1(X)$ and $\alpha_0(X)$ are replaced with random variables α_1 and α_0 (having distributions that could depend on X), provided that $(\alpha_0, \alpha_1) \perp (U, Y_1, Y_0) \mid X$.

Our results also immediately extend to permit estimation of quantile treatment effects. The proof of Theorem 2.3.2 shows that the first term in equation (2.3.1) equals $E(Y_1 \mid X)$ and the second term equals $E(Y_0 \mid X)$. Suppose we strengthen the assumption that $E(Y_j \mid X, V, U) = E(Y_j \mid X, U)$ for $j = 0, 1$ to say that $F_j(Y_j \mid X, V, U) = F_j(Y_j \mid X, U)$, where F_j is the distribution function of Y_j for $j = 0, 1$. Then one can apply Theorem 2.3.2 replacing Y with $I(Y \leq y)$ for any y, and thereby estimate $E(I(Y_j \leq y) \mid X) = F_j(y \mid X)$. Given this identification and associated estimators for the distributions $F_j(y \mid X)$ of the counterfactuals Y_j, we could then immediately recover quantile treatment effects.

2.3.3 Panel Data

We now consider a panel data version of the model, allowing for fixed effects. Let the model of treatment be

$$D_{it} = I(\alpha_0(x_{it}) \leq a_i + b_t + V_{it} + U_{it} \leq \alpha_1(x_{it})), \tag{2.3.2}$$

and let the outcome equation be

$$Y_{it} = \widetilde{a}_i + \widetilde{b}_t + Y_{0it} + (Y_{1it} - Y_{0it})D_{it}, \tag{2.3.3}$$

where a_i and $\widetilde{a}_i$ equal the coefficients of individual i dummy variables, and where b_t and $\widetilde{b}_t$ equal the coefficients time dummies in the two equations. For example, b_t is the coefficient of a dummy variable that equals one for all observations in time period t and zero otherwise.

As before, the observables in the model are the outcome Y, treatment D, instrument V, and covariate vector X. We assume that a_i, b_t, $\widetilde{a}_i$, and $\widetilde{b}_t$ for all i and t are random variables, in that we make some mild assumptions regarding their distribution. However, we interpret a_i, b_t, $\widetilde{a}_i$, and $\widetilde{b}_t$ as fixed effects, in that their values will not be estimated, their distribution is not be parameterized or estimated, and they are permitted to correlate with both X and with the unobservables in the model in unknown ways.

Assumption 22 *For all individuals i and time periods t, $a_i, b_t, \widetilde{a}_i, \widetilde{b}_t$ are random variables.*

$$E\left(\widetilde{a}_i + \widetilde{b}_t + Y_{jit} | X_{it}, V_{it}, a_i, b_t, U_{it}\right) = E\left(\widetilde{a}_i + \widetilde{b}_t + Y_{jit} \middle| X_{it}, a_i, b_t, U_{it}\right),$$

for $j = 0, 1$. $V_{it} \perp a_i, b_t, U_{it} | X_{it}$.

Remark 2.3.3 The identification permits having a_i, b_t, $\widetilde{a}_i$, *and* $\widetilde{b}_t$ be pre-determined constants.[3] We more generally let a_i, b_t, $\widetilde{a}_i$, and $\widetilde{b}_t$ for all i and t be random variables (which can be correlated with X_{it}) to clarify the minimum restrictions we require of them, which is the above conditional independence with V_{it}. Note that the joint distribution of $(a_i, b_t, \widetilde{a}_i, \widetilde{b}_t, U_{it}, Y_{0it}, Y_{1it})$ conditional or unconditional on X_{it}, is unknown. A similar assumption regarding fixed effects in discrete choice panel models appears in Honore and Lewbel (2002).

Assumption 23 *Assumption 20 holds after replacing supp$[\alpha_0(X) - U, \alpha_1(X) - U]$ with*

[3] We thank a referee for pointing this out.

supp$[\alpha_0(x_{it}) - \widetilde{a}_i - \widetilde{b}_t - U_{it}, \alpha_1(x_{it}) - \widetilde{a}_i - \widetilde{b}_t - U_{it}]$. *We similary define* $I_\tau(v_{it}, x_{it})$ *and let* $I_{\tau it} \equiv I_\tau(v_{it}, x_{it})$.

Assumptions 22 and 23 are essentially the panel data versions of Assumptions 19 and 20.

Theorem 2.3.4 *Let Assumption 18, 21, 22, and 23 hold for each individual* i *in each time period* t. *Let* f_{v_t} *denote the density of* V *in time* t. *Then*

$$\frac{E[I_{\tau it} D_{it} Y_{it} / f_{v_t}(V_{it}|X_{it})|X_{it}]}{E[I_{\tau it} D_{it} / f_{v_t}(V_{it}|X_{it})|X_{it}]} - \frac{E[I_{\tau it}(1 - D_{it}) Y_{it} / f_{v_t}(V_{it}|X_{it})|X_{it}]}{E[I_{\tau it}(1 - D_{it}) / f_{v_t}(V_{it}|X_{it})|X_{it}]} = E(Y_{1it} - Y_{0it}|X_{it}). \tag{2.3.4}$$

This theorem is proved in Appendix C. Analogous to Theorem 2.3.2, identification is also possible without the trimming $I_{\tau it}$.

In typical panel data models, removing individual specific fixed effects requires some type of differencing over time, and similarly for removing time fixed effects. Moreover, in nonlinear models such differencing is generally not possible and fixed effects need to be estimated, leading to the incidental parameters problem. However, despite the presence of fixed effects in both the linear outcome equation (2.3.3) and the nonlinear treatment equation (2.3.2), we have that equation (2.3.4) is virtually the same as the expression for $\psi(X)$ in equation (2.3.1). As a result, no differencing or incidental parameter estimation is required. The estimator for panel data, corresponding to equation (2.3.4) in Theorem 2.3.4 is essentially identical to the cross section estimator $\widehat{\psi}(x)$ based on Theorem 2.3.2.

The intuition for this result is that the same density weighting that eliminates the confounding effects of U in the cross section also happens to remove the nonlinear treatment

equation fixed effects, and the differencing of the two terms that appear in equation (2.3.4) removes the linear outcome equation fixed effects.

As in the cross section case, estimation based on equation (2.3.4) simply replaces f_{v_t} with a kernel estimator of this density, and replaces the expectations with averages, or nonparametric regressions if elements of X_{it} are continuous. If the distribution of V varies by time then the density of f_{v_t} must be estimated separately in each time period, but averaging or nonparametric regressions is done across all individuals in all time periods. No differencing or other techniques for removing the fixed effects are required.

Identification and estimation based on more general panel models is possible. We present one such extension, allowing for dynamic effects, in Appendix B.

2.3.4 Asymptotic Normality

Our identification theorems permit fixed trimming, indexed by $I_{\tau i}$ in the cross section and $I_{\tau it}$ in the panel. This trimming allows our limiting distribution derivation to follow standard arguments like those in Newey and McFadden (1994), avoiding the complications associated with kernel estimator bias when V is near the boundary of its support. As a result, we can estimate $\psi(X)$ at the standard nonparametric rate associated with the dimension of X. As noted briefly in Lewbel (2000b) and discussed more thoroughly in Khan and Tamer (2010), without fixed trimming obtaining standard convergence rates with inverse density weighted estimators like ours would generally require V to have very thick tails. Our fixed trimming avoids these issues.

For this section, standard assumptions regarding kernels, bandwidths and smoothness, as well as detailed proofs, are provided in Appendix C. Assumptions that require some discussion are kept in the main text.

Cross Section Asymptotics

We first derive properties for the cross section version of our estimator. Let x be an interior point in the support of X. Define

$$h_{1i} \equiv \frac{D_i I_{\tau i} Y_i}{f(v_i|x_i)},\ g_{1i} \equiv \frac{D_i I_{\tau i}}{f(v_i|x_i)},\ h_{2i} \equiv \frac{(1-D_i) I_{\tau i} Y_i}{f(v_i|x_i)},\ g_{2i} \equiv \frac{(1-D_i)\, I_{\tau i}}{f(v_i|x_i)},\ \psi_1(x) \equiv \frac{E\left(h_{1i}|\, x\right)}{E\left(g_{1i}|\, x\right)},\ \psi_2(x) \equiv \frac{E\left(h_{2i}|\, x\right)}{E\left(g_{2i}|\, x\right)}$$

From the proof of Theorem 2.3.2, $\psi_1(x) = E(Y_1|x)$ and $\psi_2(x) = E(Y_0|x)$. We let the sample counterpart estimator of $\psi(x) = \psi_1(x) - \psi_2(x)$ be

$$\widehat{\psi}_1(x) - \widehat{\psi}_2(x) = \frac{\frac{1}{nh^k}\sum_{i=1}^{n} \frac{D_i I_{\tau i} Y_i}{\widehat{f}(v_i|x_i)} K\left(\frac{x_i - x}{h}\right)}{\frac{1}{nh^k}\sum_{i=1}^{n} \frac{D_i I_{\tau i}}{\widehat{f}(v_i|x_i)} K\left(\frac{x_i - x}{h}\right)} - \frac{\frac{1}{nh^k}\sum_{i=1}^{n} \frac{(1-D_i) I_{\tau i} Y_i}{\widehat{f}(v_i|x_i)} K\left(\frac{x_i - x}{h}\right)}{\frac{1}{nh^k}\sum_{i=1}^{n} \frac{(1-D_i) I_{\tau i}}{\widehat{f}(v_i|x_i)} K\left(\frac{x_i - x}{h}\right)}, \tag{2.3.5}$$

where $\widehat{f}(v_i|x_i) = \widehat{f}_{xv}(x_i, v_i)/\widehat{f}_x(x_i)$ with $\widehat{f}_x(x_i)$ and $\widehat{f}_{xv}(x_i, v_i)$ being the standard leave-one-out nonparametric density estimators

$$\begin{aligned} \widehat{f}_x(x_i) &= \frac{1}{nh^k} \sum_{l=1, l\neq i}^{n} K\left(\frac{x_l - x_i}{h}\right), \\ \widehat{f}_{xv}(x_i, v_i) &= \frac{1}{nh^{k+1}} \sum_{l=1, l\neq i}^{n} K\left(\frac{x_l - x_i}{h}, \frac{v_l - v_i}{h}\right), \end{aligned}$$

where K is a kernel function and h is the bandwidth.

Assumptions 35, 36, 37 and 38 provided in Appendix C, are all standard. Given these assumptions, the asymptotic normality of estimator (2.3.5) is established as follows.

Theorem 2.3.5 *Let Assumption 18 ∼ 21, 35 ∼ 38 hold. As* $n \to \infty$, $h \to 0$, $nh^k \to \infty$,

and $nh^k h^{2p} \to c_0 \in (0, +\infty)$. *For any interior point* x *in the support of* X, *we have*

$$\frac{\sqrt{nh^k}}{var\left(q_i\left(x\right)|x\right)\int_{\mathbb{R}^k} K^2\left(u\right)du}\left[\widehat{\psi}_1(x) - \widehat{\psi}_2(x) - E\left(Y_1 - Y_0|x\right) - \mathbb{B}_p\left(x\right)\right] \xrightarrow{d} N\left(0,1\right),$$

where $q_i\left(x\right)$ *and* $\mathbb{B}_p\left(x\right)$ *are defined in equation (2.10.6) and (2.10.7) respectively in the supplemental Appendix.*

The proof is in the supplemental online appendix.

Remark 2.3.6 The unconditional treatment effect $E\left(Y_1 - Y_0\right)$ could be estimated as $\frac{1}{n}\sum_{i=1}^{n}\left[\widehat{\psi}_1(x_i) - \widehat{\psi}_2(x_i)\right]$. It is generally possible to attain parametric convergence rates for estimators like this (averages of smooth functions of kernel estimated densities and regressions), though doing so requires dealing with standard boundary bias issues for values of x near the boundary of its support. One method for doing so would be to use boundary bias corrections as in Hickman and Hubbard (2014). Another approach is to employ asymptotic trimming as in Robinson (1988) or Hardle and Stoker (1989).

Panel Data Asymptotics

The panel version of our estimator is essentially identical to averaging our cross section estimator across multiple time periods, because, as noted in the proof of Theorem 2.3.4, the estimator automatically accounts for fixed effects. Deriving the asymptotic properties of the panel estimator is therefore relatively straightforward but tedious. The main difference from the cross section case comes from allowing the distribution of V to vary over time. However, it is also necessary to keep track of the fixed effects, since they can affect the limiting distribution of the estimator.

To simplify the analysis and to focus on the new issues raised by panel data, assume we have no covariates X. This will be the case for our empirical application. Equations (2.3.2) and (2.3.3) then simplify to

$$Y_{it} = a_i + b_t + Y_{0it} + (Y_{1it} - Y_{0it})\, D_{it}, \tag{2.3.6}$$

$$D_{it} = I\left[0 \leq \widetilde{a}_i + \widetilde{b}_t + V_{it} + U_{it} \leq \alpha\right], \tag{2.3.7}$$

where $i = 1, 2, ..., n$, $t = 1, 2, ..., T$, and α is an unknown constant. The sample counterpart we estimate is then

$$\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{D_{it}I_{\tau it}Y_{it}}{\widehat{f}_{v_t}(v_{it})}}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{D_{it}I_{\tau it}}{\widehat{f}_{v_t}(v_{it})}} - \frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{(1-D_{it})I_{\tau it}Y_{it}}{\widehat{f}_{v_t}(v_{it})}}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{(1-D_{it})I_{\tau it}}{\widehat{f}_{v_t}(v_{it})}}. \tag{2.3.8}$$

If we did have covariates X_{it}, the estimator would then be analogous to equation (2.3.5), and we would need to combine the asymptotics we do here with those of the previous section.

We consider asymptotics where n goes to infinity faster than T, and obtain a convergence rate of $\sqrt{nT}$. Define ε_{jit} by $Y_{jit} = E(Y_j) + \varepsilon_{jit}$ *for* $j = 0, 1$, where $E(\varepsilon_{jit}) = 0$. Define

$$\Lambda_{1it} \equiv \frac{\left(Y_{it} - \mathrm{E}(\widetilde{a}_i + \widetilde{b}_t + Y_1)\right) D_{it}I_{\tau it} - E\left[\left(Y_{it} - \mathrm{E}(\widetilde{a}_i + \widetilde{b}_t + Y_1)\right) D_{it}I_{\tau it}\middle| v_{it}\right]}{f_{v_t}(v_{it})},$$

$$\Lambda_{2it} \equiv \frac{\left(Y_{it} - \mathrm{E}(\widetilde{a}_i + \widetilde{b}_t + Y_0)\right)(1 - D_{it})\, I_{\tau it} - E\left[\left(Y_{it} - \mathrm{E}(\widetilde{a}_i + \widetilde{b}_t + Y_0)\right)(1 - D_{it})\, I_{\tau it}\middle| v_{it}\right]}{f_{v_t}(v_{it})},$$

$$\Pi_{1it} \equiv \frac{D_{it}I_{\tau it}}{f_{v_t}(v_{it})},\ \overline{\Pi}_1 \equiv E\left(\frac{D_{it}I_{\tau it}}{f_{v_t}(v_{it})}\right),\ \Pi_{2it} \equiv \frac{(1 - D_{it})\, I_{\tau it}}{f_{v_t}(v_{it})},\ \overline{\Pi}_2 \equiv E\left(\frac{(1 - D_{it})I_{\tau it}}{f_{v_t}(v_{it})}\right).$$

Assumption 24 *$n \to \infty, T \to \infty$, and $T = o\left(n^{1-c_T}\right)$, for some $c_T \in (0, 1)$.*

Because $\sqrt{n}$ convergence of $\widehat{f}_{v_t}$ is not attainable, we need $T = o\left(n^{1-c_T}\right)$ to attain the convergence rate $\left(\widehat{f}_{v_t}(v) - f_{v_t}(v)\right)^2 = o_p\left((nT)^{-1/2}\right)$ with appropriate choice of bandwidth and kernel function.

Assumption 25 *$a_i, \widetilde{a}_i$ are i.i.d. across i and $b_t, \widetilde{b}_t$ are i.i.d. across t. (Y_{0it}, Y_{1it}) are identically distributed across i, t. $(U_{it}, Y_{0it}, Y_{1it}) \perp (U_{i't'}, Y_{0i't'}, Y_{1i't'})$ for any $i \neq i'$, $t \neq t'$. $(U_{it}, Y_{0it}, Y_{1it}) \perp (U_{it'}, Y_{0it'}, Y_{1it'}) \,|a_i, \widetilde{a}_i$ for any i, $t \neq t'$. $(U_{it}, Y_{0it}, Y_{1it}) \perp (U_{i't}, Y_{0i't}, Y_{1i't})| \, b_t, \widetilde{b}_t$ for any t, $i \neq i'$.*

The assumption that (Y_{0it}, Y_{1it}) is identically distributed over t as well as over i for each t is made only for convenience, and could be relaxed at the expense of additional notation that would include redefining the estimand to be the average value over time of $E\left(Y_1 - Y_0|t\right)$. We could allow heterogeneity (non-identical distributions) over the time dimension for other variables as well, but we do exploit the i.i.d. assumption across i, conditional on t. These i.i.d. assumptions could also be relaxed to allow for weak dependence, at the cost of requiring more notation and a more general central limit theorem. Variables with the same i or the same t subscript are correlated with each other through individual or time dummies.

In Assumption 25, we define $a_i, \widetilde{a}_i, b_t, \widetilde{b}_t$ as random variables, but we estimate the model treating them as one would handle fixed effects, without estimating their values or their distributions and without imposing the kinds of assumptions that would be required for random effects estimation. For example, a_i and b_t are allowed to be correlated with U_{it} and Y_{it} in arbitrary unknown ways.

Remark 2.3.7 Although they are not estimated, $\widetilde{a}_i$ and $\widetilde{b}_t$ do affect our limiting distribution, because the weights on these variables in the first and second components of our

estimator are not identical in finite samples. In Lemma 2.10.8 in the online supplemental appendix, we show that the difference in these components due to $\widetilde{a}_i$ and $\widetilde{b}_t$ is $O_P\left((nT)^{-1/2}\right)$.

Assumption 26 *V_{it} are independent across i and t. V_{it} are identically distributed across i given t, with distribution $f_{v_t}(V_{it})$.*

For each time period t, Assumption 26 is equivalent to the cross section special regressor assumption without X. In addition it is assumed that special regressor observations are independent over time, but the distribution of V_{it} is allowed to vary with t. This independence assumption could be relaxed, and it would even be possible to let V_{it} be fixed over time for each i, though this would require dropping the cross section fixed effects from the model.

Assumption 27 $E\left(\varepsilon_{0it}|a_i,\widetilde{a}_i\right) = E\left(\varepsilon_{1it}|a_i,\widetilde{a}_i\right)$ *and* $E\left(\varepsilon_{0it}|b_t,\widetilde{b}_t\right) = E\left(\varepsilon_{1it}|b_t,\widetilde{b}_t\right)$.

This is assumption is somewhat stronger than the assumption needed to interpret ATE, because we only need $E\left(\varepsilon_{jit}\right) = 0$ such that $E\left(Y_{jit}\right) = E\left(Y_j\right)$ for $j = 0, 1$.

Remark 2.3.8 Assumption 27 is necessary to attain $\sqrt{nT}$-convergence. To see why the assumption is necessary, suppose we could observe the counterfactuals Y_{1it} and Y_{0it}. Then the direct estimator for $E\left(Y_1\right) - E\left(Y_0\right)$ would just be $\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\left(Y_{1it} - Y_{0it}\right)$. The random component for this estimator is $\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\left(\varepsilon_{1it} - \varepsilon_{0it}\right)$, which is equal to

$$\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\left(\varepsilon_{1it} - \varepsilon_{0it} - E\left(\varepsilon_{1it} - \varepsilon_{0it}|a_i,\widetilde{a}_i\right) - E\left(\varepsilon_{1it} - \varepsilon_{0it}|b_t,\widetilde{b}_t\right)\right)$$
$$+\frac{1}{n}\sum_{i=1}^{n}E\left(\varepsilon_{1it} - \varepsilon_{0it}|a_i,\widetilde{a}_i\right) + \frac{1}{T}\sum_{t=1}^{T}E\left(\varepsilon_{1it} - \varepsilon_{0it}|b_t,\widetilde{b}_t\right).$$

The first term is $O_P\left((nT)^{-1/2}\right)$, the second term is $O_P\left(n^{-1/2}\right)$, and the third term is $O_P\left(T^{-1/2}\right)$. So the convergence rate of this estimator is $O_P\left(T^{-1/2}\right)$ if $E\left[E\left(\varepsilon_{1it}-\varepsilon_{0it}|b_t,\widetilde{b}_t\right)^2\right] > 0$. So even in the infeasible case where counterfactuals are observable, Assumption 27 would be necessary to obtain $\sqrt{nT}$-convergence instead of rate $\sqrt{T}$.

As was discussed earlier, if potential outcomes are given by $\widetilde{Y}_{dit} = \widetilde{a}_i + \widetilde{b}_t + Y_{dit}$, then our estimand $E\left(Y_{1it} - Y_{0it}\right)$ equals a standard ATE $E\left(\widetilde{Y}_{1it} - \widetilde{Y}_{0it}\right)$.

Additional Assumptions 37 and 39 provided in the Appendix are standard. Given these assumptions, the rate $\sqrt{nT}$ asymptotic normality of estimator (2.3.8) is established as follows.

Theorem 2.3.9 *Let Assumption 18, 21, 22, 23, 24, 25, 26, 27, 37, 39 hold. Assume that bandwidth $h = c_0 n^{-c_T/2}$ in $\widehat{f}_{v_t}$, and assume a kernel of order $p \geq (1 - c_T/2)/c_T$. Then*

$$\begin{aligned}
&\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} D_{it}Y_{it}\Big/\widehat{f}_{v_t}(v_{it})}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} D_{it}\Big/\widehat{f}_{v_t}(v_{it})} - \frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}(1-D_{it})Y_{it}\Big/\widehat{f}_{v_t}(v_{it})}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}(1-D_{it})\Big/\widehat{f}_{v_t}(v_{it})} - [E(Y_1) - E(Y_0)] \\
= \;& \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\left(\frac{\Lambda_{1it}}{\overline{\Pi}_1} - \frac{\Lambda_{2it}}{\overline{\Pi}_2}\right) + o_P\left((nT)^{-1/2}\right),
\end{aligned}$$

and $\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\left(\frac{\Lambda_{1it}}{\overline{\Pi}_1} - \frac{\Lambda_{2it}}{\overline{\Pi}_2}\right) = O_p\left((nT)^{-1/2}\right)$.

Remark 2.3.10 This theorem gives the influence function $\frac{\Lambda_{1it}}{\overline{\Pi}_1} - \frac{\Lambda_{2it}}{\overline{\Pi}_2}$ for our estimator. The terms in the influence function are identically distributed. From Lemma 2.10.8, 2.10.9, and 2.10.10 in the supplemental appendix, those terms are dependent (through fixed effects) but not correlated with each other. Additional assumptions on the dependence of those terms are needed to establish asymptotic normality.

Remark 2.3.11 Suppose $\left(a_i, \widetilde{a}_i, b_t, \widetilde{b}_t\right)$ is a series of constants instead of random variables. From the proof of Lemma 2.10.8, our estimator will still be consistent as long as $\frac{1}{n^2T}\left(\sum_{i=1}^{n}\widetilde{a}_i^2\right) = o(1)$ and $\frac{1}{nT^2}\left(\sum_{t=1}^{T}\widetilde{b}_t^2\right) = o(1)$. The estimator will also, given Assumption 27, still converge at rate $\sqrt{nT}$ with the same limiting distribution given below if $\frac{1}{n}\left(\sum_{i=1}^{n}\widetilde{a}_i^2\right) = O(1)$ and $\frac{1}{T}\left(\sum_{t=1}^{T}\widetilde{b}_t^2\right) = O(1)$. This result allows for limited forms of time trends of unknown form, e.g., b_t and $\widetilde{b}_t$ could systematically increase or decrease over time.

Some additional results involving panel data asymptotics are provided in the appendix. In particular, we provide limiting distribution theory under some more general conditions, including if Assumption 27 does not hold, and a more general model of fixed effects.

2.4 Competition and Innovation

We apply our model to test the the "Inverted-U" theory of ABBGH (Aghion, Bloom, Blundell, Griffith, and Howitt 2005) relating innovation investments to competitiveness in an industry. ABBGH consider two types of oligopoly industries, called Neck-and-Neck (NN) industries, in which firms are technologically close to equal, and Leader-Laggard (LL) industries, where one firm is technologically ahead of others. For these industries there are two opposing effects of competition on innovation. One is the *Schumpeterian effect*, where increased competition reduces profits and thus reduces the incentive to innovate. The second is the *escape-competition effect,* where firms innovate to increase the profits associated with being a leader. For these latter firms, increased competition increases the incentive to innovate. ABBGH argue that the escape-competition effect dominates in NN industries while the Schumpeterian effect dominates in LL industries. This theory results

in an inverted-U relationship, because low levels of competition are associated with NN industries and hence with low innovation, by the escape-competition effect, and high levels of competition are associated with LL industries, again leading to low innovation but now by the Schumpeterian effect. In contrast, with an intermediate level of competition, both NN and LL industries innovate to some extent, yielding a higher overall level of innovation in steady state than in either the low or high competition industries.

ABBGH find empirical support for the inverted-U based mainly on UK data. Hashmi (2013) revisits the relationship using a richer dataset from the US, and finds no inverted-U. Hashmi notes that his finding can be reconciled with the ABBGH model by the assumption that the manufacturing industries in the UK are, on the average, more neck and neck than their counterparts in the US.

For identification and estimation, both the ABBGH and Hashmi empirical results depend heavily on functional form assumptions, by fully parameterizing both the relationship of competitiveness to innovation and the functional form of error distributions. In contrast, we apply our model to test for an inverted-U relationship with minimal restrictions on functional forms and error distributions.

2.4.1 Data

Our sample, from Hashmi (2013), consists of US three-digit level industry annual data from 1976 to 2001. There are 116 industries, resulting in 2716 industry-year observations. Our analysis is based on three key variables: a measure of industry competitiveness, a measure of industry innovation, and a source-weighted average of industry exchange rates that serves as an instrument, and hence as our special regressor. Summary statistics for this data are reported in Table 1. We only applied our estimator to Hashmi's data and not to ABBGH's

data, because the latter does not contain a continuous instrumental variable that can be used as a special regressor.

The measure of the level of competition for industry i at time t, denoted c_{it}, is defined by

$$c_{it} = 1 - \frac{1}{n_{it}} \sum_{j=1}^{n_{it}} l_{jt}, \tag{2.4.1}$$

where i indexes firms, l_{jt} is the Lerner index of the price the cost margin of firm j in year t, and n_{it} is the number of firms in industry i in year t. The higher c_{it} is, the higher is the level of competition. The innovation index, denoted y_{it}, is a measure of citation-weighted patent counts, constructed using data from the NBER Patent Data Project. Details regarding the construction of this data can be found in Hashmi (2013).

As ABBGH point out, innovation and competition are endogenous, that is, there are likely to exist unobserved characteristics of each industry i in each time period t that can affect both. To deal with this endogeneity, Hashmi uses a source-weighted average of industry exchange rates as instrument variable for competition (ABBGH use a different, events related instrument). Hashmi's instrument, V_{it}, is a weighted average of the US dollar exchange rate with the currencies of trading partners, with weights that vary by industry according to the share of each country in the imports to the US. This instrument has been used in other similar applications, including Revenga (1990, 1992) and Bertrand (2004).

2.4.2 Model Specifications

Hashmi (2013) adopts a control function approach to deal with endogeneity. In a first stage, c_{it} is regressed on V_{it}, industry dummies and time dummies, so

$$c_{it} = V_{it}\beta + a_i + b_t + w_{it}, \tag{2.4.2}$$

where a_i and b_t are fixed effects (coefficients of industry and time dummies) and w_{it} is the error from the first stage regression. The fitted residuals $\widehat{w}_{it}$ from this regression are then included as additional regressors in an outcome equation of the form

$$\ln(y_{it}) = \widetilde{a}_i + \widetilde{b}_t + \theta_0 + \theta_1 c_{it} + \theta_2 c_{it}^2 + \delta\widehat{w}_{it} + \varepsilon_{it}, \tag{2.4.3}$$

where $\widetilde{a}_i$ and $\widetilde{b}_t$ are outcome equation fixed effects (coefficients of industry and time dummies). Hashmi estimates the coefficients in equation (2.4.3) by maximum likelihood, where the distribution of errors ε_{it} is determined by assuming that y_{it} has a negative binomial distribution, conditional on c_{it}, industry, and year dummies. This model assumes the relationship of $\ln(y)$ to c is quadratic, with an inverted-U shape if θ_1 is positive and θ_2 negative. The industry and time dummies cannot be differenced out in this model, and so are estimated along with the other parameters.

In addition to the possibility that this quadratic is misspecified, or that the endogeneity takes a form that is not completely eliminated by the control function addition of $\widehat{w}$ as a regressor, or that the distribution is not negative binomial, Hashmi's estimates could also suffer the from the problem of incidental parameters (Neyman and Scott 1948). This problem is that the need to estimate industry and time fixed effects results in inconsistent

parameter estimates unless both T and n go to infinity. In this application neither T nor n is particularly small, but the presence of the fixed effects still results in over 100 nuisance parameters to estimate, which can lead to imprecision. Our intention is not to criticize Hashmi's or ABBGH's model, but only to point out that there are many reasons why it is desirable to provide a less parametric alternative, to verify that their results are not due to potential model specification or estimation problems.

To apply our estimator, let the treatment indicator D_{it} equal one for industries i that have neither very low nor very high levels of competition in period t, and otherwise let $D_{it} = 0$. We then let innovation y_{it} be determined by

$$y_{it} = \widetilde{a}_i + \widetilde{b}_t + Y_{0it} + (Y_{1it} - Y_{0it})D_{it}. \tag{2.4.4}$$

where $\widetilde{a}_i$, $\widetilde{b}_t$ are the industry and time dummies respectively, and Y_{0it} are Y_{1it} are unobserved potential outcomes for industry i in time t, after controlling for time and industry fixed effects. Unlike the error distribution imposed in equation (2.4.3), both Y_{1it} and Y_{0it} here are random variables with completely unknown distributions that can be correlated with each other, and with the error term in the D_{it} equation, in completely unknown ways. We will then estimate the ATE $E(Y_{1it}-Y_{0it})$, which equals the average difference in outcomes y (after controlling for fixed effects), between industries with moderate levels of competitiveness, versus industries that have very low or very high levels of competitiveness.

What our model assumes about the treatment indicator D_{it} is

$$D_{it} = I(\alpha_0 \leq a_i + b_t + V_{it} + U_{it} \leq \alpha_1), \tag{2.4.5}$$

where a_i and b_t are industry and time dummies, U_{it} are unobserved, unknown factors that affect competition, and α_0 and α_1 are unknown constants. The way to interpret equation (2.4.5) is that the latent variable c^*_{it} given by

$$c^*_{it} = a_i + b_t + V_{it} + U_{it} \tag{2.4.6}$$

is some unobserved true level of competitiveness of industry i in time t. Our model does not require the observed competitiveness measure c_{it} to equal the true measure c^*_{it}, but if they do happen to be equal then our model is consistent with having Hashmi's equation (2.4.2) hold. Note when comparing the models for c^*_{it} and c_{it} to each other that replacing c^*_{it} with βc^*_{it} to make equation (2.4.6) line up with equation (2.4.5) is a free scale normalization that can be made without loss of generality, because the definition of D_{it} is unaffected by rescaling c^*_{it}.[4]

As in Hashmi's model, our estimator assumes that V_{it} is a valid instrument, affecting competitiveness c^*_{it} and hence the treatment indicator D_{it}, but not directly affecting the outcome y_{it}. We also require that V_{it} has a large support. This appears to be the case in our data, e.g., the exchange rate measure sometimes as much as doubles or halves over time even within a single industry, and varies substantially across industries as well.

2.4.3 Measurement Errors in Competitiveness

In our empirical application, we define D_{it} to be one when the observed c_{it} lies between the .25 and .75 quantiles of the empirical c_{it} distribution (we also experiment with other

[4] In our data it is very unlikely that c_{it} perfectly measures true competitiveness in each industry and time period. However, if c_{it} is not mismeasured, then the thresholds used to construct D_{it} from c_{it} would be proportional, up to the scaling of the coefficient of V, to the unknown thresholds $\alpha_1(X)$ and $\alpha_0(X)$ (after accounting for unknown fixed effects a_i and b_t). In theory this information might be usable to increase estimation efficiency, by exploiting the fact that $E(D/f_v|X)$, which we estimate, equals $\alpha_1(X) - \alpha_0(X)$.

quantiles). This is therefore consistent with equation (2.4.2) if c_{it} is linear in c^*_{it}. However, our model remains consistent even if c_{it} differs greatly from c^*_{it}, as long as the middle 50% of industry and time periods in the c_{it} distribution corresponds to the middle 50% of industry and time periods in the c^*_{it} distribution.

More generally, suppose c_{it} equals c^*_{it} plus some measurement error. Then the Hashmi model, even if correctly specified, will be consistent only if this measurement error satisfies the conditions necessary for validity of their control function estimator. Some control function estimators remain consistent in models containing measurement errors that are classical, i.e., independent of the true c^*_{it} and of the true model. However, the Hashmi control function estimator would not be consistent even with classical measurement errors, because equation (2.4.3) is nonlinear in the potentially mismeasured variable c_{it} (this is not intended as a criticism of Hashmi's empirical application, since that work uses control functions only to deal with endogeneity and never made any claims regarding measurement errors).

In contrast, our estimator can remain consistent in theory even with measurement errors that are large and nonclassical, as long as c_{it} correctly sorts industries into moderate versus non-moderate levels of competitiveness. However, in practice, measurement error in c_{it} will likely cause some industries to be misclassified, so D_{it} is likely to be mismeasured for some industries (particularly for some that are near the .25 and .75 quantile cutoffs). Also, in practice we should expect Hashmi's control function specification to at least partly correct for potential measurement error.

To summarize: competitiveness is difficult to precisely define and measure, and as a result the impact of measurement errors on this analysis could be large. One advantage

of our methodology is that it only depends on sorting industries into two groups (that is, moderate versus extreme levels of competitiveness as indicated by D_{it}). While this sorting discards some information and may therefore cost some efficiency, it will also mitigate measurement error biases, because only a small number of observations of D_{it} are likely to be mismeasured even if most or all of the c_{it} observations are mismeasured to some extent. To check whether this intuition is correct, in an appendix we do a monte carlo analysis that compares the accuracy of our estimator with that of Hashmi's in the presence of measurement errors.

2.4.4 Estimation

Our estimator is quite easy to implement, in part because it does not entail any numerical searches or maximizations. We first estimate the density of V_{it} separately for each year, using a standard kernel density estimator $\widehat{f}_{v_t}(v_{it}) = \frac{1}{n-1}\sum_{j\neq i,,j=1}^{n} \frac{1}{h} K\left(\frac{v_{it}-v_{jt}}{h}\right)$. Note that the density is estimated at each of the data points v_{it}. We employ a Gaussian kernel function K, and choose the bandwidth h using Silverman's rule of thumb. Our estimator involves dividing by these nonparametric density estimates, which can result in outlier observations when $\widehat{f}$ is close to zero. As suggested in Lewbel (2000a) and Dong and Lewbel (2015) for other special regressor based estimators, we trim out (i.e., discard from the sample) the 2% of observations with the smallest values of $\widehat{f}_{v_t}$. This defines the trimming function $I_\tau(v)$ from our asymptotic theory.

Given the density estimates $\widehat{f}_{v_t}(v_{it})$, our resulting estimate of the ATE $E(Y_{1it} - Y_{0it})$ is

then given by

$$\text{Trim-ATE} = \frac{\sum_i \sum_t I_\tau(v_{it}) D_{it} Y_{it} / \widehat{f}_{v_t}(V_{it})}{\sum_i \sum_t I_\tau(v_{it}) D_{it} / \widehat{f}_{v_t}(V_{it})} - \frac{\sum_i \sum_t I_\tau(v_{it})(1 - D_{it}) Y_{it} / \widehat{f}_{v_t}(V_{it})}{\sum_i \sum_t I_\tau(v_{it})(1 - D_{it}) / \widehat{f}_{v_t}(V_{it})} \quad (2.4.7)$$

where the i and t sums are over the 98% of observations that were not trimmed out. This model corresponds to the estimator (2.3.8), which has standard errors that we calculate based on the asymptotic distribution provided in Theorem 2.8.2. To assess the effect of the trimming on this estimator, we construct a corresponding estimate of ATE that is not trimmed, given by

$$\text{No-Trim-ATE} = \frac{\sum_i \sum_t D_{it} Y_{it} / \widehat{f}_{v_t}(V_{it})}{\sum_i \sum_t D_{it} / \widehat{f}_{v_t}(V_{it})} - \frac{\sum_i \sum_t (1 - D_{it}) Y_{it} / \widehat{f}_{v_t}(V_{it})}{\sum_i \sum_t (1 - D_{it}) / \widehat{f}_{v_t}(V_{it})}. \quad (2.4.8)$$

For comparison, in addition we calculate a Naive-ATE estimator given by

$$\text{Naive-ATE} = \frac{\sum_i \sum_t D_{it} Y_{it}}{\sum_i \sum_t D_{it}} - \frac{\sum_i \sum_t (1 - D_{it}) Y_{it}}{\sum_i \sum_t (1 - D_{it})}. \quad (2.4.9)$$

This Naive-ATE just subtracts the average value of Y_{it} when $D_{it} = 0$ from the average value of Y_{it} when $D_{it} = 1$. This would be a consistent estimator of the ATE if treatment were unconfounded, that is, if low or high competitiveness as indicated by D_{it} was randomly assigned over firms and time periods. One could also consider a LATE estimator such as an instrumental variables regression of Y on D using V as an instrument. However, as noted in the introduction, LATE requires that the probability of treatment increase monotonically with the instrument. This requirement does not hold in our application, since both increasing or decreasing V sufficiently causes the probability of treatment to

decrease.

We also compare our results to a parametric maximum likelihood estimate of the ATE (denoted ML-ATE) assuming a Heckman (1979) type selection model for treatment. This model assumes equations (2.4.4) and (2.4.5) hold and that U, Y_0, Y_1 are jointly normally distributed. Let Φ denote the standard normal cumulative distribution function, $\theta_0 = E(Y_0)$, $\theta_1 = E(Y_1)$, and $\sigma = cov[U, Y_0, Y_1]$ be the three by three covariance matrix of elements σ_{kl} for $k = 1, 2, 3$ and $l = 1, 2, 3$. Then the ML-ATE is defined by

$$\text{ML-ATE} = \widehat{\theta}_1 - \widehat{\theta}_0 \quad \text{where} \quad \left[\widehat{\theta}_0, \widehat{\theta}_1, \widehat{\alpha}_0, \widehat{\alpha}_1, [\widehat{\sigma}_{kl}]_{3\times 3}\right] = \arg\max \sum_i \sum_t$$
$$\left\{ (1 - D_{it}) \log \left(\frac{1}{\sigma_{22}} \phi \left(\frac{Y_{it} - \theta_0}{\sigma_{22}} \right) \left[\Phi \left(\frac{\alpha_0 - V_{it} - \frac{\sigma_{12}}{\sigma_{22}} (Y_{it} - \theta_0)}{\sqrt{\sigma_{11} - \sigma_{12}^2/\sigma_{22}}} \right) + 1 - \Phi \left(\frac{\alpha_1 - V_{it} - \frac{\sigma_{12}}{\sigma_{22}} (Y_{it} - \theta_0)}{\sqrt{\sigma_{11} - \sigma_{12}^2/\sigma_{22}}} \right) \right] \right) \right.$$
$$\left. + D_{it} \log \left(\frac{1}{\sigma_{33}} \phi \left(\frac{Y_{it} - \theta_1}{\sigma_{33}} \right) \left[\Phi \left(\frac{\alpha_1 - V_{it} - \frac{\sigma_{13}}{\sigma_{33}} (Y_{it} - \theta_1)}{\sqrt{\sigma_{11} - \sigma_{13}^2/\sigma_{33}}} \right) - \Phi \left(\frac{\alpha_0 - V_{it} - \frac{\sigma_{13}}{\sigma_{33}} (Y_{it} - \theta_1)}{\sqrt{\sigma_{11} - \sigma_{13}^2/\sigma_{33}}} \right) \right] \right) \right\}.$$

2.4.5 Empirical Results

Figure 1 shows our kernel density estimates $\widehat{f}_{v_t}$ for each year t. The estimates can be seen to vary quite a bit over time, so we use separate density estimates for each year instead of assuming a constant distribution across years. Figure 2 shows a scatterplot of our competitiveness and innovation data. The fitted linear line using least squared errors estimation is slightly downward sloping. The fitted quadratic line using least squared errors estimation is slightly U-shape. Note that the fitted curves do not deal with the endogeneity issue.

Table 2A shows our main empirical results. The first row of Table 2A provides estimates where D_{it} is defined to equal one for the middle half of the data, that is, D_{it} equals one for firms and years that lie between the 25th and 75th percent quantiles of the observed

measure of competition, making half the observations treated and the other half untreated. Other rows of Table 2A report results using different quantiles to define D_{it}. In each row of Table 2A we report four estimates of ATE, as described in the previous section. Standard errors for all the estimates are provided in parentheses.

An inverted-U would imply a positive ATE, but all of our estimates are negative, confirming Hashmi's finding that the inverted-U is not present in US data. For example, our main estimate from the first row of Table 2A is that the Trim-ATE equals -3.9, and is strongly statistically significant. We also find that failure to appropriately control for error correlations between competitiveness and innovation substantially biases the magnitudes of estimated treatment effects. Our semiparametric estimates of the ATE are 50% to 100% larger than both the naive estimates that ignore these correlations, and the maximum likelihood estimates that allow for correlations but requires the errors to be jointly normally distributed.

Attempts to find a positive ATE by experimenting with more unusual quantiles for defining D_{it} were for the most part fruitless. An exception, based on examination of Figure 2, was to define the left and right thresholds by 0.62 (10%) and 0.68 (20%) respectively. This implies a heavily skewed inverted U where 80% of firms are in the upper tail. This yields a positive ATE of 8.66, but this model is implausible, since it treats a very narrow spike in Figure 2 as the set of all moderately sized firms. We also experimented with varying the degree of trimming, but we only report results without trimming and with 2% percent trimming because the impacts of other changes in trimming were very small.

The quantiles of c_{it} vary over time, so instead of defining D_{it} based on quantiles of the entire sample of c_{it} observations, one could instead define D_{it} for each year t based on the

quantiles of c_{it} just in year t. As a robustness check, results are reported in Table 2B based on estimates calculating D_{it} this alternative way. Comparing Table 2A and 2B, shows that the results are quite similar using either definition.

Hashmi models the mean of innovation using equation (2.4.3), so the following object constructed from Hashmi's paper can be compared with our ATE estimates

$$E\left(\exp\left(\widetilde{a}_i+\widetilde{b}_t+\theta_0+\theta_1 c_{it}+\theta_2 c_{it}^2+\delta w_{it}\right)\Big|\, D_{it}=1\right)-E\left(\exp\left(\widetilde{a}_i+\widetilde{b}_t+\theta_0+\theta_1 c_{it}+\theta_2 c_{it}^2+\delta w_{it}\right)\Big|\, D_{it}=0\right).$$

We estimate this by replacing the expectations with D_{it} cell means, and using Hashmi's estimates for the parameters $\widetilde{a}_i$, $\widetilde{b}_t$, θ_0, θ_1, θ_2, and δ.[5] The value of this quantity we find from his model is -1.8, which is about half of our estimated ATE and similar to the ML-ATE and Naive-ATE. Again, we agree with Hashmi's main result regarding signs of effects, but not magnitudes. This discrepancy might come from misspecification of Hashmi's model, sensitivity to measurement error in c_{it} in his model, or imprecision in his estimates of $\widetilde{a}$ and $\widetilde{b}$ due to the incidental parameters problem.

We provide a further comparison of our results with Hashmi's in Section 2.5.2, where we extend our identification result to the full ordered choice model of treatment.

2.4.6 Monte Carlo Designed for the Empirical Example

To assess how the estimator works in small samples, we provide two sets of Monte Carlo experiments. We designed these experiments to closely match moments and other features of our empirical data, to see how likely our estimator is to perform well in a controlled setting that mimics our actual application. The number of observations is set to 2716, the

[5] Hashmi only reports θ_1 and θ_2. These are in the first column of table 9 in Hashmi (2013). Other parameter estimates are found in the Stata log file he posts online.

same as the number of observations in our empirical dataset. The same four estimators we applied on the actual data, Trim-ATE, No-Trim-ATE, Naive-ATE and ML-ATE, are analyzed in each set of Monte Carlo simulations

Let e_{1i}, e_{2i}, e_{3i}, and V_i be random variables that are drawn independently of each other. We consider a few different distributions for these variables as described below. The counterfactual outcomes in our simulation are defined by

$$Y_{0i} = \theta_0 + \theta_{01} e_{1i} + \theta_{02} e_{3i} \text{ and } \; Y_{1i} = \theta_1 + \theta_{11} e_{2i} + \theta_{12} e_{3i}.$$

True competitiveness is constructed to equal $V_i + \theta_2 e_{3i}$, and treatment D_i is defined to equal one for observations i that lie between the 25th and 75th quantile of the distribution of $V_i + \theta_2 e_{3i}$. The observed outcome is then constructed as

$$Y_i = Y_{0i} + (Y_{1i} - Y_{0i}) D_i.$$

For simplicity, fixed effect type dummies are omitted from the model. Note that e_{3i} appears in D_i, Y_{0i}, and Y_{1i}, and so is the source of confounding in this model. By construction, the unobserved U_i in our theoretical model is given by $U_i = \theta_2 e_{3i}$. Let θ denote the vector of parameters $(\theta_0, \theta_1, \theta_2, \theta_{01}, \theta_{02}, \theta_{11}, \theta_{12})$. In each Monte Carlo experiment the parameter vector θ is set to match moments and outcomes of our actual data, specifically, they are set to make the ATE $\theta_1 - \theta_0$ equal our estimate -3.90, and to make the mean and variance of Y_i and D_i, and the covariance between Y_i and D_i, equal the values observed in our data. The variance of V_i is freely normalized (inside the binomial response indicator) to equal one.

The ML-ATE estimator is asymptotically efficient when e_{1i}, e_{2i}, and e_{3i} are normally distributed. In our first experiment we let e_{1i}, e_{2i}, e_{3i}, and V_i each have a standard normal distribution, so the resulting ML-ATE estimates can then serve as an efficient benchmark.

As noted by Khan and Tamer (2010), single threshold crossing model special regressor estimators converge at slow rates when f_v has thin tails, as in the previous design. Although their results are not directly applicable to this paper's two threshold model, it is still sensible to see if our estimator works better with thicker tails, so our second experiment gives e_{1i}, e_{2i}, e_{3i}, and V_i each a uniform distribution on $[-0.5, 0.5]$. Note this is still likely not the best case for our estimator, since Khan and Tamer (2010) note that special regressor methods converge fastest when V has a thick tail and all other variables have thin tails.

Both the normal and uniform designs have symmetric errors, which favors the ML alternative over our estimator. However, with symmetric errors it is impossible to define a vector θ that matches all the moments of the empirical data, because symmetry prevents matching the empirical covariance between Y and D. Therefore, in both designs we choose values for θ that match all the other moments and come as close as possible to matching this covariance (the required values for θ are given in the footnote of Table 4).

To match the empirical correlation between Y and D along with other moments, we next consider designs that introduce asymmetry into the confounder e_{3i}. In our third experiment, we let e_{1i}, e_{2i}, and V_i be standard normal and let e_{3i} be a modified normal, equaling a standard normal with probability one half when $e_{3i} < 0$ and equaling θ_3 times a standard normal with probability one half when $e_{3i} \geq 0$. When then choose θ_3 along with the other elements of θ to match the moments of the empirical data including the covariance of Y with D. This required setting $\theta_3 = 2.65$. Similarly, in a fourth experiment

we let e_{1i}, e_{2i}, and V_i be uniform on $[-0.5, 0.5]$ and take e_{3i} to equal a (demeaned) mixed uniform distribution. This mixture was uniform on $[-2, 0]$ with probability one half and uniform on $[0, 5]$ with probability one half, before demeaning.

Each of these four Monte Carlo experiments was replicated 10,000 times, and the results are summarized in Table 4 in the supplemental Appendix. Panel A in Table 4 is the symmetric normal design. Because of symmetry, all of the estimators in this design are unbiased. ML, being efficient here, has the lowest root mean squared error (RMSE), and the naive estimator is almost as efficient as ML in this case, since it just involves differencing simple covariance estimates. Our Trim-ATE estimator performs reasonably well compared to the efficient estimator, being unbiased and having a RMSE of .43 versus the efficient .30. Trimming improves the RMSE enormously here, as expected because f_v has thin tails, which produces outliers in the denominator of averages weighted by f_v.

Panel B of Table 4 shows that, in the symmetric uniform design, all four estimators are almost identical. The happens because, with V is uniform, $\widehat{f_v}$ is close to a constant, and the estimators for the average effects of the treated and the untreated are close to their sample means.

In the asymmetric designs, given in panels C and D of Table 4, the ML-ATE and Naive-ATE are no longer consistent, and both become substantially downward biased, with an average value of about one half the true value of -3.90. In contrast, our trimmed and untrimmed ATE estimates had far smaller downward biases, resulting in much smaller RMSE, particularly for the Trim-ATE.

The differences in biases between the inconsistent estimators (ML-ATE and NAIVE-ATE) and our proposed estimator in these asymmetric Monte Carlos closely match the

observed differences in our empirical application estimates. Specifically, in case 1 of Table 2A the estimated ATE using the ML and Naive estimators is about one half the estimate of -3.90 we obtained using Trim-ATE. This provides evidence that the Monte Carlo results in panels C and D of Table 4 are relevant for assessing the empirical performance of our proposed estimator.

In addition to assessing the quality of estimators we also assess the quality of associated standard error estimates, by providing, in the last column of Table 4, the percentage of times the true ATE fell in the estimated 95% confidence interval (defined as the estimated ATE plus or minus two estimated standard errors). In the symmetric designs all the estimated standard errors for all the estimators were too large, yielding overly conservative inference. In the asymmetric designs the estimated 95% confidence intervals of the inconsistent estimators ML-ATE and NAIVE-ATE were very poor, containing the true value less than 25% of the time. The No-Trim-ATE did much better, but our preferred estimator, Trim-ATE, was by far the best, giving correct 95% coverage in panel C, and conservative 99% coverage in panel D.

2.5 Extensions

2.5.1 Testing the Large Support Assumption

The large support assumption is crucial for identification. In this section, we provide a formal test on this assumption.

Suppose $\text{supp}(V) = [-M', M]$, $M', M > 0$, f_v bounded away from zero, and the support of U is also a fixed interval on the real line. For simplicity, we assume there is no covariates

X and

$$D = I\left(0 \leq V + U \leq \alpha\right). \tag{2.5.1}$$

The large support assumption of V in the paper is that $\text{supp}(-U) \subseteq \text{supp}(V)$, $\text{supp}(\alpha - U) \subseteq \text{supp}(V)$. Under the model specification and that the supports of U and V are both fixed intervals on the real line, the large support assumption holds if and only if $P\left(D = 1|V = M\right) = 0$ and $P\left(D = 1|V = -M'\right) = 0$.

Without loss of generality, we only discuss how we test $P\left(D = 1|V = M\right) = 0$, and the test for the other part of the implications follows similarly.

As discussed in the Theorem 2.5.3 and Remark 2.5.4, when the support of V strictly covers the support of $\alpha - U$ on the right end, the test statistic will degenerate to a constant 0. When the support of V is exactly the same as the support of $\alpha - U$ on the right end, the test statistic will converge at $\sqrt{n}$ rate, even though we estimate $P\left(D = 1|V = M\right)$ nonparametrically. Though this property looks nice, it is basically not helpful on inference and thus on test. Because of this peculiar property, we decide to compromise a bit and instead test

$$\mathbb{H}_0 : P\left(D = 1|V = M\right) \geq \varepsilon^*,$$

where ε^* is a pre-determined small value.

Suppose we have n observations, we let $v_n^{(1)} = \max\left\{v_i, i = 1, ..., n\right\}$. Then by Lemma 2.10.14, $M - v_n^{(1)} = O_P\left(n^{-1}\right)$. We approximate M by $\widehat{M} = v_n^{(1)}$. Denote $G_D\left(v\right) \equiv P\left(D = 1|V = v\right)$. We let $G'_{D,-}\left(M\right)$ and $G''_{D,-}\left(M\right)$ denote the left first and second derivatives at M respectively. Since we are only interested in the estimation at the boundary, we estimate it by the local linear regression, which is known to be able to correct boundary

effects automatically:

$$\min_{\boldsymbol{\beta}} \frac{1}{n}\sum_{i=1}^{n}\left(I\left(D_i=1\right)-\beta_0-\beta_1\left(\frac{V_i-v}{h}\right)\right)K_h\left(V_i-v\right),$$

where $\boldsymbol{\beta}\equiv\left(\beta_0,\beta_1\right)^T$, $K_h\left(V_i-v\right)\equiv\frac{1}{h}K\left(\frac{V_i-v}{h}\right)$. Let $\widehat{\boldsymbol{\beta}}\left(v\right)\equiv\left(\widehat{\beta}_0\left(v\right),\widehat{\beta}_1\left(v\right)\right)^T$ be the estimates from the above estimation.

Let $e_1=\left(1,0\right)^T$. We are only interested in the estimates at the boundary point M. Since we do not know M, we approximate it by $\widehat{M}$. Re-arrange the data such that $v_n=v_n^{(1)}=\widehat{M}$. Then it could be obtained from the following leave-one-out estimator $\widehat{G}_D\left(\widehat{M}\right)\equiv e_1^T\widehat{\beta}\left(\widehat{M}\right)$ where

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}\left(\widehat{M}\right) &= \left[\mathbf{S}_h\left(\widehat{M}\right)\right]^{-1}\frac{1}{n-1}\sum_{i=1}^{n-1}K_h\left(V_i-\widehat{M}\right)\begin{pmatrix}1\\ \left(V_i-\widehat{M}\right)/h\end{pmatrix}I\left(D_i=1\right) \qquad (2.5.2)\\
\mathbf{S}_h\left(\widehat{M}\right) &\equiv \frac{1}{n-1}\sum_{i=1}^{n-1}K_h\left(V_i-\widehat{M}\right)\begin{pmatrix}1\\ \left(V_i-\widehat{M}\right)/h\end{pmatrix}\left(1,\left(V_i-\widehat{M}\right)/h\right).
\end{aligned}$$

Define $S_{j,-}\equiv\int_{-\infty}^{0}K\left(u\right)u^j du$ for any positive integer j, and $\overline{\mathbf{S}}\equiv\begin{pmatrix}S_{0,-} & S_{1,-}\\ S_{1,-} & S_{2,-}\end{pmatrix}$. Not hard to see that $\overline{\mathbf{S}}$ is the limit of $\mathbf{S}_h\left(\widehat{M}\right)$ in probability.

Theorem 2.5.1 *Under $\mathbb{H}_0$, i.i.d., the model specification (2.5.1), that $G_D\left(v\right)$ is twice differentiable and $h=c_0n^{-1/5}$ for some $c_0>0$, we have*

$$\sqrt{nh}\left(\widehat{G}_D\left(\widehat{M}\right)-G_D\left(M\right)-\mathbb{B}_h\right)\xrightarrow{d}N\left(0,\sigma^2\left(M\right)\right),$$

$$where\, \mathbb{B}_h \equiv e_1^T \overline{\mathbf{S}}^{-1} \begin{pmatrix} S_{2,-} \\ S_{3,-} \end{pmatrix} G''_{D,-}(M) f_v(M) h^2, \sigma^2(M) \equiv e_1^T \overline{\mathbf{S}}^{-1} \mathbf{Q} \overline{\mathbf{S}}^{-1} e_1 G_D(M)(1 - G_D(M)) f_v(M),$$

$$\mathbf{Q} \equiv \begin{pmatrix} Q_{0,-} & Q_{1,-} \\ Q_{1,-} & Q_{2,-} \end{pmatrix}, Q_{j,-} \equiv \int_{-\infty}^{0} K^2(u) u^j du.$$

The theorem is proved in the supplemental appendix.

By Theorem 2.5.1, we can test $\mathbb{H}_0$ via standard z-test. The P-value is calculated as $P = \Phi\left(\frac{\widehat{G}_D(\widehat{M}) - \widehat{\mathbb{B}}_h - \varepsilon^*}{\widehat{\sigma}(M)}\right)$, where $\widehat{\mathbb{B}}_h, \widehat{\sigma}(M)$ can be obtained as in Remark 2.5.2. We suggest $\varepsilon^* = 0.05$ in empirical applications.

Remark 2.5.2 Not hard to see that the optimal bandwidth is

$$h_{\text{opt}} = n^{-1/5} \left[\left(e_1^T \overline{\mathbf{S}}^{-1} \mathbf{Q} \overline{\mathbf{S}}^{-1} e_1 G_D(M)(1 - G_D(M)) f_v(M) \right) \Big/ \left(e_1^T \overline{\mathbf{S}}^{-1} \begin{pmatrix} S_{2,-} \\ S_{3,-} \end{pmatrix} G''_{D,-}(M) f_v(M) \right)^2 \right]^{1/5}.$$

To get $\widehat{\mathbb{B}}_h$, $\widehat{\sigma}^2(M)$ and h_{opt}, one can estimate $G''_{D,-}(M)$ and $f_v(M)$ separately using local quadratic estimator and the kernel estimator with boundary correction (e.g., Hardle 1990)[6], respectively. Fan and Gijbels (1992) has discussed the nice property and the nice performance of this plug-in estimator.

To make this section more complete, below we list the asymptotic property of $\widehat{G}_D\left(\widehat{M}\right)$ when the support of U covers the support of $\alpha - U$ on the right end.

Theorem 2.5.3 *Under i.i.d., the model specification (2.5.1), that $G_D(v)$ is twice differentiable, that the support of U covers the support of $\alpha - U$ on the right end, and $h = c_0 n^{-2/5}$,*

[6] Here we use the modified kernel function: $K(x) \Big/ \int_{-\infty}^{0} K(x)\, dx$.

for some $c_0 > 0$, *we have*

$$\sqrt{n}\left(\widehat{G}_D\left(\widehat{M}\right) - G_D\left(M\right)\right) \xrightarrow{d} N\left(0, \widetilde{\sigma}^2\left(M\right)\right),$$

where $\widetilde{\sigma}^2\left(M\right) \equiv e_1^T \overline{\mathbf{S}}^{-1} \mathbf{Q} \overline{\mathbf{S}}^{-1} e_1 G_D'\left(M\right) f_v\left(M\right)$, $\mathbf{Q} \equiv \begin{pmatrix} Q_{1,-} & Q_{2,-} \\ Q_{2,-} & Q_{3,-} \end{pmatrix}$, $Q_{j,-} \equiv \int_{-\infty}^{0} K^2\left(u\right) u^j du$.

The theorem is proved in the supplemental appendix.

Remark 2.5.4 In the case where the support of V strictly covers the support of U, $G_D\left(v\right) = G_D'\left(v\right) = 0$ in an interval around the boundary point. In this case, according to the above theorem, $\sigma^2\left(M\right) = 0$. The estimates $\widehat{G}_D\left(\widehat{M}\right)$ will degenerate to zero in the limit.

We conduct this test of the large support assumption for our data by testing $\mathbb{H}_0$: $P\left(D = 1 | V = M\right) \geq \varepsilon^*$ where we set ε^* as 0.05. We use the result from Theorem 2.5.1. The P-value is calculated as $P = \Phi\left(\frac{\widehat{G}_D\left(\widehat{M}\right) - \widehat{\mathbb{B}}_h - \varepsilon^*}{\widehat{\sigma}(M)}\right)$ as in Remark 2.5.2.

We first ignore the fixed effects and use the whole data set to get the estimates of $\widehat{G}_D$ at the left and right boundary (minimum and maximum of V respectively). We use the optimal bandwidth as in Remark 2.5.2. It turns out that $\widehat{G}_D$ at both sides are all very close to zero. The P-values are both 0.000, which reject the null hypothesis that $P\left(D = 1 | V = M\right) \geq 0.05$.

If V_{it} and U_{it} are homogenous in terms of the support across different time periods, we are all set on the testing. To be safe, we also do the test for each time period separately. We need to do the test at both ends of V for 26 periods. Each period contains at most 116 observations. The results are collected in Table 3A. The null hypothesis is rejected at 5% significance level in 36 out of the whole 52 cases (both sides for 26 years); the null is

rejected at 10% significant level in 37 out of the whole 52 cases. The results are not perfect. However, we think at least some of the failures are probably due to the small sample size.

One feature of the tests is that the P-value can easily become 0. The intuition can be seen from the theoretical property of the test: when the support of V strictly covers the support of U, the estimate should be close to zero and the variance is also every small (see Remark 2.5.4), which will drive $P = \Phi\left(\frac{\widehat{G}_D(\widehat{M}) - \widehat{\mathbb{B}}_h - \varepsilon^*}{\widehat{\sigma}(M)}\right)$ to zero.

To summarize, the null hypothesis is rejected for the whole sample in favor of our large support assumption. The null hypothesis has been rejected for most cases when we conduct the test over each time period separately. We think we could conclude that the large support assumption generally holds for this application.

2.5.2 Ordered Choice Identification at Infinity

We now consider an extension of our results to full ordered choice data. In our empirical application, this extension will help us distinguish between competing alternatives to the inverted-U hypothesis.

We change notation in this section to define three values for treatment and three corresponding potential outcomes: Let $D = 0$ with potential outcome Y_0 if the latent variable is below the lower threshold, $D = 1$ with potential outcome Y_1 if the latent variable is between the two thresholds, and $D = 2$ with Y_2 if above the upper threshold. In this notation, our previous results did not distinguish observing $D = 0$ from $D = 2$. In practice, one would usually be able to see if an individual had $D = 0$ vs $D = 2$. Following Heckman, Urzua and Vytlacil (2006) (see also Andrews and Schafgans 1998), if one can distinguish $D = 0$ from $D = 2$, then one could use identification at infinity to identify $E(Y_0|X)$ and $E(Y_2|X)$.

Suppose the treatment indicator is the standard ordered choice as follows

$$D = I\left[\alpha_0\left(X\right) \leq V + U < \alpha_1\left(X\right)\right] + 2I\left[V + U \geq \alpha_1\left(X\right)\right]. \tag{2.5.3}$$

The outcome equation is

$$Y = Y_0 I\left(D = 0\right) + Y_1 I\left(D = 1\right) + Y_2 I\left(D = 2\right). \tag{2.5.4}$$

As noted by Heckman, Urzua and Vytlacil (2006), without invoking functional form assumptions, identification of $E\left(Y_0|X\right)$ and $E\left(Y_2|X\right)$ requires $E\left(D|X,V\right) \to 0$ and $E\left(D|X,V\right) \to 2$ for limiting values of one or more covariates. In our case we use this identification at infinity technique to identify $E\left(Y_0|X\right)$ taking the limit of $E\left(D|X,V\right)$ as V gets sufficiently small, and identifying $E\left(Y_2|X\right)$ by taking the limit as V gets sufficiently large.

We impose the following assumptions, which are very similar to our earlier assumptions, except now we assume all three values of D are observable.

Assumption 28 *We observe realizations of an outcome Y, multinomial treatment indicator D taking three values 0 1 and 2, a covariate V, and a $k \times 1$ covariate vector X. Assume the outcome Y and treatment indicator D are given by equations (2.5.4) and (2.5.3) respectively, where $\alpha_0\left(X\right)$ and $\alpha_1\left(X\right)$ are unknown threshold functions and $\alpha_0\left(X\right) < \alpha_1\left(X\right)$, U is an unobserved latent random error, and Y_0 Y_1 Y_2 are unobserved random potential outcomes for $D = 0, 1, 2$ respectively. The joint distribution of $\left(U, Y_0, Y_1, Y_2\right)$, either unconditional or conditional on X, is unknown.*

Assumption 29 *Assume $V \perp \left(U, Y_0, Y_1, Y_2\right) \mid X$. For all $x \in supp\left(X\right)$, the $supp(V \mid X = x)$ covers the $supp\left(\alpha_1\left(X\right) - U\right)$ on the right and covers $supp(\alpha_0\left(X\right) - U \mid X = x)$ on the*

left.

Theorem 2.5.5 (**Identification**) *Suppose Assumption 28, 29 hold.* $\{\gamma_n(X)\}_{n=1}^{\infty}, \{\gamma'_n(X)\}_{n=1}^{\infty}$ *are increasing series such that* $\lim_{n\to\infty} E(D \mid X, V \leq -\gamma_n(X)) = 0$ *and* $\lim_{n\to\infty} E(D \mid X, V \geq \gamma'_n(X)) = 2$. *Then*

$$E(Y_0 \mid X) = \lim_{n\to\infty} E(I(D=0)Y \mid X, V \leq -\gamma_n(X)), \quad E(Y_2 \mid X) = \lim_{n\to\infty} E\left(I(D=2)Y \mid X, V \geq \gamma'_n(X)\right).$$

$$\text{and} \quad E(Y_1 \mid X) = \frac{E[I(D=1)Y/f(V \mid X) \mid X]}{E[I(D=1)/f(V \mid X) \mid X]}/$$

The proof is in the supplemental Appendix.

The tuning paramerers $\gamma_n(X)$ and $\gamma'_n(X)$ determine the set of V values that we average over as the sample size grows. The intuition of this identification at infinity is that the the larger in magnitude are $\gamma_n(X)$ and $\gamma'_n(X)$, the more extreme are the values of V that we average over, and hence the lower is the probability that the confounder U can alter D. Eventually, the effect of the confounder vanishes in the limit.

In the special case of our empirical example, the treatment indicator is defined as

$$D_{it} = I[\alpha_0 \leq V_{it} + a_i + b_t + U_{it} < \alpha_1] + 2I[V_{it} + a_i + b_t + U_{it} \geq \alpha_1], \tag{2.5.5}$$

where covariates X become a constant, and the former confounder becomes $a_i + b_t + U_{it}$. For the outcome equation, compared with the previous sections, $\widetilde{a}_i, \widetilde{b}_t$ are now absorbed into the outcome variable.

The sample counterpart estimators for $E(Y_0)$, $E(Y_1)$ and $E(Y_2)$ (corresponding to

$E\left(\widetilde{a}_i + \widetilde{b}_t + Y_j\right)$, $j = 1, 2, 3$, in our previous notation) based on the above theorem is

$$\widehat{E}(Y_0) = \frac{\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} I(D_{it} = 0) Y_{it} I(V_{it} \leq -\gamma_{nT})}{\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} I(D_{it} = 0) I(V_{it} \leq -\gamma_{nT})}, \widehat{E}(Y_2) = \frac{\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} I(D_{it} = 2) Y_{it} I(V_{it} > \gamma'_{nT})}{\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} I(D_{it} = 2) I\left(V_{it} > \gamma'_{nT}\right)},$$

$$\widehat{E}(Y_1) = \frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} I(D_{it} = 1) Y_{it} \Big/ \widehat{f}_{v_t}(v_{it})}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} I(D_{it} = 1) \Big/ \widehat{f}_{v_t}(v_{it})}$$

where γ_{nT} and γ'_{nT} are increasing series such that $\lim_{n,T\to\infty} E(D \mid V \leq -\gamma_{nT}) = 0$ and $\lim_{n,T\to\infty} E(D \mid V \geq \gamma'_{nT}) = 2$.

The estimator $\widehat{E}(Y_1)$ duplicates our previous results, based on Theorem 2.3.4. Andrews and Schafgans (1998) and Schafgans (1998) provide asymptotics for models that are identified at infinity like $\widehat{E}(Y_0)$ and $\widehat{E}(Y_2)$. Following Schafgans (1998), we consider various values for γ_{nT} and γ'_{nT}, based on the percentage of uncensored observations (e.g., for $E(Y_0)$ uncensored means $D = 0$) used in the estimation, specifically, 50%, 40%, 30%, 20%, 10%, and 5%. We estimate a standard error for $\widehat{E}(Y_0)$ using

$$\frac{\left(\sum_{i=1}^{n}\sum_{t=1}^{T} I(D_{it} = 0)\left(Y_{it} - \widehat{E}(Y_0)\right)^2 I(V_{it} \leq -\gamma_{nT})\right)^{1/2}}{\sum_{i=1}^{n}\sum_{t=1}^{T} I(D_{it} = 0) I(V_{it} \leq -\gamma_{nT})}.$$

and similarly for $\widehat{E}(Y_2)$.

Empirical results are reported in Table 3B. Panel A displays the estimates when we define D as a whole sample and the thresholds are the 25% and 75% quantiles of the measure of the competitiveness. As a robustness check, panel B displays the results when

we define D separately for each year. The results do not vary much by year, and are also not very sensitive to the choice of γ_{nT} and γ'_{nT}, especially for $\widehat{E}(Y_2)$. Not surprisingly, the standard errors become larger when the tuning parameters are larger, corresponding to averages over fewer observations.

The estimate $\widehat{E}(Y_1)$ from the previous section is 4.33. Seen from Table 3B, $\widehat{E}(Y_2)$ is slightly (but not significantly) higher than $\widehat{E}(Y_1)$, while $\widehat{E}(Y_0)$ is much higher than $\widehat{E}(Y_1)$ and $\widehat{E}(Y_2)$. Therefore we obtain a mostly decreasing relationship between innovation and competition. This pattern is similar to the quadratic least squares estimation of the raw data (see Figure 2). This result is also consistent with Hashmi (2013), who speculates that manufacturing in the US is probably dominated by Leader-Laggard industries.

The results in the section come with some caveats that do not apply to our main identification theorem. Due to only being identified at infinity, the estimators $\widehat{E}(Y_0)$ and $\widehat{E}(Y_2)$ will converge slower than the parametric rate. These estimates can also be sensitive to the tuning parameters γ_{nT} and γ'_{nT}.

The results here could be readily extended to cases having more than three outcomes. For example, if we had four outcomes, we can identify middle outcomes involving $E(Y_1)$ and $E(Y_2)$ by the special regressor approach using Theorem 2.3.2, and those outcomes at the ends, i.e. $E(Y_0)$ and $E(Y_4)$, using identification at infinity.

2.6 Conclusions

In this article, we propose a new method to estimate the average treatment effect in a two threshold model, where the treated group is a middle choice. In our application, treatment is defined as facing an intermediate level of competition, versus a low or high level of

competition.

The proposed model is confounded, because the unobservables that affect the treatment indicator D can be correlated in unknown ways with potential outcomes Y_0 and Y_1, with or without conditioning on other covariates. No parametric or semiparametric restrictions are placed on distributions of treatment and potential outcomes, so treatment effects are not identified by functional form. Our model assumes a continuous instrument V with large support, but treatment effects are not identified at infinity, because both very large and very small values of V drive the probability of treatment close to zero, while no value of V (or of other covariates) drives the probability of treatment close to one. So in this framework none of the conditions that are known to permit point identification of the ATE hold. Even the monotonicity conditions usually required for identifying LATE are not satisfied. Nevertheless, we show that the ATE is identified, using a special regressor argument, and we provide conditions under which the corresponding estimate of the ATE is consistent, and asymptotically normal. Root-n consistency is even obtained in a panel context with fixed effects, despite nonlinearities that would usually induce an incidental parameters problem in the equation determining probability of treatment. We provide Monte Carlo results that show that our estimator works well in small samples (comparable to the data in our empirical application). We show in an Appendix that our estimator is relatively robust to measurement error and misspecification.

We use our method to investigate the relationship between competition and innovation. Our estimates using a dataset from Hashmi (2013) confirm Hashmi's findings that an inverted-U is not present in US data. We also find that standard parametric model and naive treatment effect estimators substantially underestimate the magnitude of the

treatment effect in this context.

Bibliography

[1] Abadie, A., J. Angrist, and G. Imbens (2002), "Instrumental Variables Estimation of Quantile Treatment Effects,"Econometrica. Vol. 70, No. 1, 91-117.

[2] Aghion, P., N. Bloom, R. Blundell, R. Griffith, and P. Howitt (2005), "Competition and Innovation: an Inverted-U Relationship," Quarterly Journal of Economics, 120(2):701-28.

[3] Andrews, D., and M. Schafgans (1998), "Semiparametric Estimation of the Intercept of a Sample Selection Model," Review of Economic Studies, 65, 497-17.

[4] Angrist, J. D., and G.W. Imbens (1995), "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity," Journal of the American Statistical Association 90:430, 431-42.

[5] Ashenfelter, O. (1978), "Estimating the Effect of Training Programs on Earnings," Review of Economics and Statistics, 60, 47-57.

[6] Ashenfelter, O., and D. Card (1985), "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs," Review of Economics and Statistics, 67, 648-660.

[7] Bai J. (2009), “Panel Data Models with Interactive Fixed Effects”, Econometrica, 77, 1229-1279.

[8] Barnow, B. S., G. G. Cain, and A. S. Goldberger (1980), “Issues in the Analysis of Selectivity Bias,” in Evaluation Studies, Vol. 5, ed. by E. Stromsdorfer and G. Farkas. San Francisco: Sage, 43–59.

[9] Bertrand, M. (2004), "From the Invisible Handshake to the Invisible Hand? How Import Competition Changes the Employment Relationship," Journal of Labor Economics, 22(4):722-65.

[10] Bitler, M., J. Gelbach, and H. Hoynes (2006), “What Mean Impacts Miss: Distributional Effects of Welfare Reform Experiments,” American Economic Review, 96, 4, 988-1012.

[11] Björklund, A., and R. Moffitt (1987), “The Estimation of Wage Gains and Welfare Gains in Self–Selection Models,” Review of Economics and Statistics, Vol. LXIX, 42–49.

[12] Cao, S., R. Moineddin, M.L. Urquia, F. Razak, and J.G. Ray (2014), "J-shapedness: an Often Missed, Often Miscalculated Relation: the Example of Weight and Mortality," Journal of Epidemiology and Community Health, 68, 683–690.

[13] Card, D. (1990), “The Impact of the Mariel Boatlift on the Miami Labor Market,” Industrial and Labor Relations Review 43, 245-257.

[14] Card, D., and A. Krueger (1993), “Trends in Relative Black-White Earnings Revisited,” American Economic Review, vol. 83, no. 2, 85-91.

[15] Card, D., and A. Krueger (1994), “MinimumWages and Employment: A Case Study of the Fast-food Industry in New Jersey and Pennsylvania,” American Economic Review, 84 (4), 772-784.

[16] Cecchetti, S. G., and E. Kharroubi (2012), "Reassessing the Impact of Finance on Growth", working paper.

[17] Chernozhukov, V., and C. Hansen (2005), “An IV Model of Quantile Treatment Effects,” Econometrica, 73(1), 245-261.

[18] Chernozhukov, V., I. Fernandez-Val, J. Hahn, and W. Newey (2009), "Identification and Estimation of Marginal Effects in Nonlinear Panel Models," Technical report, CEMMAP.

[19] Cook, P.J., and G. Tauchen (1982), "The effect of Liquor Taxes on Heavy Drinking," Bell Journal of Economics, 13(2): 379-90.

[20] Cook, P.J., and G. Tauchen (1984), "The Effect of Minimum Drinking age Legislation on Youthful Auto Fatalities, 1970-1977," Journal of Legal Studies, 13(1): 169-90.

[21] Cochran, W., and D. Rubin (1973), “Controlling Bias in Observational Studies: A Review,” Sankhyā, 35, 417–446.

[22] Dong Y., and A. Lewbel (2015), “A Simple Estimator for Binary Choice Models with Endogenous Regressors,” Econometric Reviews, 34, 82-105.

[23] Fan, J., and I. Gijbels (1992), "Variable Bandwidth and Local Linear Regression Smoothers," The Annals of Statistics, 20(4), 2008-2036.

[24] Firpo, S. (2006), “Efficient Semiparametric Estimation of Quantile Treatment Effects,” Econometrica, 75(1), 259-276.

[25] Hardle, W. (1990), Applied Nonparametric Regression. Cambridge University Press.

[26] Hardle, W., and T. M. Stoker (1989), "Investigating Smooth Multiple Regression by the Method of Average Derivatives," Journal of the American Statistical Association, 84, 986-995

[27] Hashmi, A.R. (2013), "Competition and Innovation: the Inverted-U Relationship Revisited," Review of Economic Statistics, 95, 5, 1653-1668.

[28] Heckman, J. (1979), "Sample Selection Bias as a Specification Error," Econometrica, 47(1), 153-162.

[29] Heckman, J. J., and S. Navarro (2007), "Dynamic Discrete Choice and Dynamic Treatment Effects," Journal of Econometrics, 136, (2), 341-396.

[30] Heckman, J., and R. Robb (1984), “Alternative Methods for Evaluating the Impact of Interventions,” Longitudinal Analysis of Labor Market Data, ed. by J. Heckman and B. Singer. Cambridge, U.K.: Cambridge University Press, 156–245.

[31] Heckman, J.J., S. Urzua, and E. Vytlacil (2006), "Understanding Instrumental Variables in Models with Essential Heterogeneity," Review of Economics and Statistics, 88(3), 389-432.

[32] Heckman, J. J., and E. Vytlacil (1999), “Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects,” Proceedings of the National Academy of Science, USA, 96, 4730–4734.

[33] Heckman, J. J., and E. Vytlacil (2005), “Structural Equations, Treatment Effects, and Econometric Policy Evaluation,” Econometrica, 73, 669–738.

[34] Heckman, J. J., and E. Vytlacil (2007a), “Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Evaluation of Public Policies,” Handbook of Econometrics, J.J. Heckman and E.E. Leamer (eds.), Vol. 6, North Holland, Chapter 70.

[35] Heckman, J. J., and E. Vytlacil (2007b), “Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast their Effects in New Environments,” Handbook of Econometrics, J.J. Heckman and E.E. Leamer (eds.), Vol. 6, North Holland, Chapter 71.

[36] Hickman, B. R., and T. P. Hubbard (2014), "Replacing Sample Trimming with Boundary Correction in Nonparametric Estimation of First-Price Auctions", Journal of Applied Econometrics, forthcoming.

[37] Honore, B., and A. Lewbel (2002), "Semiparametric Binary Choice Panel Data Models Without Strictly Exogenous Regressors," Econometrica, 70, 2053-2063.

[38] Huang, J. (2015), "Banking and Shadow Banking," working paper.

[39] Imbens, G., and J. Angrist (1994), “Identification and Estimation of Local Average Treatment Effects,” Econometrica, Vol. 61, No. 2, 467-476.

[40] Imbens, G., and J. Wooldridge (2009), "Recent Development in the Econometrics of Program Evaluation," Journal of Economic Literature, 47:1, 5-86.

[41] Khan, S., and E. Tamer (2010), "Irregular Identification, Support Conditions, and Inverse Weight Estimation," Econometrica, 6, 2021-2042.

[42] Kitagawa, T. (2009), "Identification Region of the Potential Outcome under Instrument Independence," working paper.

[43] Koppes, L.L., J.M. Dekker, H.F. Hendriks, L.M. Bouter, and R.J. Heine (2005), "Moderate Alcohol Consumption Lowers the Risk of Type 2 Diabetes: a Meta–Analysis of Prospective Observational Studies," Diabetes Care, 28, 719–25.

[44] Lewbel, A. (1998), "Semiparametric Latent Variable Model Estimation with Endogenous or Mismeasured Regressors," Econometrica, 66, 105-122.

[45] Lewbel, A. (2000a), "Semiparametric Qualitative Response Model Estimation with Unknown Heteroscedasticity and Instrumental Variables," Journal of Econometrics, 97, 145-177.

[46] Lewbel, A. (2000b), "Asymptotic Trimming for Bounded Density Plug-in Estimators," working paper.

[47] Lewbel, A. (2007), "Endogenous Selection or Treatment Model Estimation," Journal of Econometrics, 141, 777-806.

[48] Lewbel, A. (2012), "An Overview of the Special Regressor Method," a chapter in Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics, Oxford University Press.

[49] Lewbel, A., Y. Dong, and T.T. Yang (2012), "Comparing Features of Convenient Estimators for Binary Choice Models With Endogenous Regressors," Canadian Journal of Economics, 45, 809-829.

[50] Manski, C. F., and J. V. Pepper (2013), "Deterrence and the Death Penalty: Partial Identification Analysis Using Repeated Cross Sections," Journal of Quantitative Criminology 29 (1), 123-141.

[51] Masry, E., 1996. Multivariate local polynomial regression for time series: uniform strong consistency rates. Journal of Time Series Analysis 17, 571-599.

[52] Meyer, B., K. Viscusi, and D. Durbin (1995), "Workers' Compensation and Injury Duration: Evidence from a Natural Experiment," American Economic Review, Vol. 85, No. 3, 322-340.

[53] Newey, W. K., and D. McFadden (1994), "Large Sample Estimation and Hypothesis Testing," in Handbook of Econometrics, vol. iv, ed. by R. F. Engle and D. L. McFadden, pp. 2111-2245, Amsterdam: Elsevier.

[54] Neyman, J., and E.L. Scott (1948), "Consistent Estimation from Partially Consistent Observations," Econometrica 16, 1-32.

[55] Revenga, A. (1990), "Essays on Labor Market Adjustment and Open Economics." PhD diss., Harvard University, Economics Department.

[56] Revenga, A. (1992), "Exporting Jobs? The Impact of Import Competition on Employment and Wages in U.S. Manufacturing," Quarterly Journal of Economics, 107, 1255–1284.

[57] Robinson, P. M. (1988), "Root-n-consistent Semiparametric Regression," Econometrica, 56, pp. 931-954.

[58] Rosenbaum, P., and D. Rubin (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," Biometrika, 70, 41–55.

[59] Rubin, D. (1974), "Estimating Causal Effects of Treatments in Randomized and Non-Randomized Studies," Journal of Educational Psychology, 66, 688–701.

[60] Schafgans, M.M.A. (1998), "Ethnic Wage Differences in Malaysia: Parametric and Semiparametric Estimation of the Chinese-Malay Wage-Gap," Journal of Applied Econometrics, 13, 481-504.

[61] Solomon C.G., F.B. Hu, M.J. Stampfer, et al. (2000), "Moderate Alcohol Consumption and Risk of Coronary Heart Disease among Women with Type 2 Diabetes Mellitus," Circulation, 102, 494–99.

2.7 Appendix A: Robustness to Measurement Errors

Observable indices of competitiveness of an industry, like the average Lerner index in equation (2.4.1), may be relatively crude measures of true competitiveness. In this section we therefore assess the robustness of our estimator, relative to a parametric model estimator like Hashmi's, to measurement error in the index of competitiveness. We first show that both models, as one would expect, become inconsistent if competitiveness is mismeasured, even when the models are otherwise correctly specified. However, we also show that the bias in our estimator resulting from measurement error is quite small relative to alternative estimators.

First consider the case where competitiveness is mismeasured, but a parametric model like Hashmi's (dropping fixed effects for simplicity) is the correct specification in terms of true competitiveness. This model assumes

$$\ln Y = \theta_0 + \theta_1 c^* + \theta_2 c^{*2} + \widetilde{e}, \tag{2.7.1}$$

where $\ln Y$ is logged innovation, c^* is the true level of competitiveness, and $\widetilde{e}$ is an error term. For simplicity we ignore discreteness in $\ln Y$, and we assume c^* can be linearly decomposed into the observable instrument V and an unobserved independent component W, so

$$c^* = V + W. \tag{2.7.2}$$

Assume validity of Hashmi's control function type assumption that $\widetilde{e} = \lambda W + e$ where e is independent of W and V, so

$$\ln Y = \theta_0 + \theta_1 c^* + \theta_2 c^{*2} + \lambda W + e \tag{2.7.3}$$

In this model, if c^* were observed, then control function estimation (first regressing c^* on a constant and V, getting the residuals $\widehat{W}$, and then regressing $\ln Y$ on a constant, c^*, c^{*2}, and $\widehat{W}$) would consistently estimate the θ coefficients and hence any desired treatment effects based on θ.

Now assume the observable competitiveness measure c equals the true measure c^* plus measurement error c_e, so

$$c = c^* + c_e, \tag{2.7.4}$$

where c_e is the measurement error and independent of c^* and e. To take the best case scenario for the parametric model, assume that the measurement error c_e has mean zero and is independent of V, W, and e.

Substituting equation (2.7.4) into equation (2.7.3) gives

$$\ln Y = \theta_0 + \theta_1 c + \theta_2 c^2 + \lambda W + e^* \tag{2.7.5}$$

where

$$e^* = \theta_1 c_e - 2\theta_2 c c_e - \theta_2 c_e^2 + e.$$

The error e^* does not have mean zero and correlates with c and c^2, which makes the control function estimator inconsistent. Unlike the case of linear models with independent mean zero measurement errors, the control function estimator is not consistent because of the nonlinearity in this model.

Now consider applying our nonparametric estimator to this model. The treatment indicator D that we would construct is defined as equaling one for firms in the .25 to .75 quantile of c and zero otherwise, while the corresponding indicator D^* based on the true measure of competitiveness equals one for firms in the .25 to .75 quantile of c^* and zero otherwise. Unless the measurement error c_e is extremely large, for the large majority of firms D will equal D^*. This is part of what makes our estimator more robust to measurement error. Even if all firms have c mismeasured to some extent, most will still be correctly classified in terms of D.

To check the relative robustness of these estimators to measurement error, we perform additional Monte Carlo analysis. As before, we construct simulated data to match moments

and the sample size of the empirical data set, and to make what would be the true treatment effect in the model match our empirical estimate of -3.9. We do two simulations, one using normal errors and one based on uniform errors, as before. In both, V and W are scaled to have equal magnitudes, so $V = \delta_0 + \delta_1\varepsilon_1$ and $W = \delta_0 + \delta_1\varepsilon_2$. To match data moments, the normal error simulations set $\delta_0 = 0.375$, $\delta_1 = 0.0733$, and $c_e = \kappa_1\varepsilon_3$ where ε_1, ε_2, and ε_3 are independent standard normals and κ_1 is a constant with values that we vary to obtain different magnitudes of measurement error. The uniform error simulations set $\delta_0 = \delta_1 = 0.25$, and $c_e \sim \kappa_2(\varepsilon_3 - 0.5)$, where now ε_1, ε_2, and ε_3 are independent random variables that are uniformly distributed on $[0, 1]$.

To check for robustness against an alternative specification as well as measurement error, we also generate data replacing the quadratic form in equation (2.7.1) with the step function

$$\ln Y = \theta_0 + (\theta_1 - \theta_0)D^* + \widetilde{e}, \tag{2.7.6}$$

where D^*, D, c^*, c, V, W, and e are all defined as above.

The Monte Carlo results, based on 10,000 replications, are reported in Tables 5 and 6 in the supplemental Appendix. In addition to trying out the four estimators we considered earlier, (Trim-ATE, No-Trim-ATE, Naive-ATE, and ML-ATE) we also apply the control function estimator described above, analogous to Hashmi's estimator.

Our main result is that, with both normal and uniform errors, the greater the magnitude of measurement error is (that is, the larger the κ_1 and κ_2 are), the better our estimator performs relative to other estimators. For the quadratic model without measurement error the control function would be a consistent parametric estimator and so should outperforms our semiparametric estimator. We find this also holds with very small measurement error

(e.g., $\kappa_1 = .02$ in the left side block of Table 5), however, both control function and Trim-ATE perform about equally at $\kappa_1 = .03$, and at the still modest measurement error level of $\kappa_2 = .04$, Trim-ATE has smaller RMSE (root mean squared error) than all the other estimators, including control function. Similar results hold for the uniform error model reported in Table 6. Also, in the step function model (shown on the right side of Tables 5 and 6) our Trim-ATE is very close to, or superior to, all the other estimators including control functions at all measurement error levels.

It is worth noting that possible measurement error affects our empirical application only because we defined treatment D in terms of an observed, possibly mismeasured underlying variable, competitiveness. In other applications the treatment indicator may be observed without error even when an underlying latent measure is completely unobserved. For example, suppose an outcome Y is determined in part by an individual's chosen education level, which in turn is determined by an ordered choice specification. The true education level of a student might be unobserved, but a treatment D defined as having graduated high school but not college could still be correctly measured.

2.8 Appendix B: Additional Extensions

2.8.1 Identifying an additive function of V

In previous sections, we assumed V appears in the selection equation in the form $V + U$. In this section, we consider the generalization where selection depends on $\varsigma(V) + U$ for some unknown function $\varsigma(V)$. This may be more realistic in some applications, since economic theory may not indicate a priori the function $\varsigma(V)$. Given identification and an associated estimator for ς, one could then redefine V as $\varsigma(V)$ and then estimate treatment effects as

before. Though not likely to be empirically relevant, it is interesting to note that in the very special case where the function ς equals the distribution function of V, the model becomes unconfounded and our proposed estimator reduces to standard propensity score weighting.

To identify ς, suppose that the selection equation takes the form

$$D = I\left(\widetilde{\alpha}_0\left(X\right) \leq \varsigma\left(V\right) + \varpi\left(X, Z\right) + U \leq \widetilde{\alpha}_1\left(X\right)\right), \tag{2.8.1}$$

for some continuously distributed exogenous covariate Z that affects selection but does not affect the thresholds. Formally, we assume the following.

Assumption 30 *Equation (2.8.1) holds for observed covariates V, Z, and vector X, where $\varsigma, \varpi, \widetilde{\alpha}_0, \widetilde{\alpha}_1$ are unknown functions, ς is differentiable, 0 is in the support of V, $\varsigma\left(0\right) = 0$, and $\varsigma'\left(0\right) = 1$, and $\left(V, Z\right) \perp U \mid X$.*

We could equivalently write equation (2.8.1) as

$$D = I\left(\alpha_0\left(X, Z\right) \leq \varsigma\left(V\right) + U \leq \alpha_1\left(X, Z\right)\right)$$

for some unknown functions ς, α_0, and α_1 where $\alpha_1\left(X, Z\right) - \alpha_0\left(X, Z\right) = \delta\left(X\right)$ for some function δ. In the standard specification of ordered choice models where $D = I\left(\delta_0 \leq X'\beta_1 + V\beta_2 + U \leq \delta_1\right)$ and X is exogenous, every continuous regressor contained in the vector X could be relabeled as Z and would then satisfy Assumption 30. This is much stronger than necessary, since we only the require existence of one such regressor.

The assumptions that zero is in the support of V, that $\varsigma\left(0\right) = 0$, and that $\varsigma'\left(0\right) = 1$ are all free normalizations that are made without loss of generality. To see this, first note that there must exist some value of v in the support of V for which $\varsigma'\left(v\right) \neq 0$, since otherwise

$\varsigma(V)$ would be a constant, not a function of V. Redefining V as $V - v$ then ensures that zero is the support of V and that $\varsigma'(0) \neq 0$. Next redefine all of the unknown functions, and U, by dividing them all by $\varsigma'(0)$. After this scale normalization, we will have by construction that $\varsigma'(0) = 1$. Finally, $\varsigma(0) = 0$ is a free location normalization, since if $\varsigma(0) = c \neq 0$ then we can redefine $\varpi(X, Z)$ as $\varpi(X, Z) + c$ to make $\varsigma(0) = 0$.

The following theorem shows identification of the function ς. The proof is constructive, so one could obtain a consistent estimator of ς by mimicking the steps of the proof, using standard kernel based nonparametric regression derivative estimators. After estimating ς, our previous estimators may be applied by replacing the density of V with the density of $\varsigma(V)$.

Theorem 2.8.1 *Suppose we observe X, V, Z, D and D follows equation (2.8.1). Given Assumption 30, the functions $\varsigma(V)$ and $\frac{\partial \varpi(X,Z)}{\partial Z}$ are identified.*

The proof is in the supplemental Appendix.

2.8.2 Additional Panel Data Asymptotics

We showed earlier that in the panel model, Assumption 27 was necessary for obtaining a $\sqrt{nT}$ convergence rate. Here we consider asymptotics when Assumption 27 is not imposed. In this case we can also replace Assumption 24 with the weaker Assumption 31, yielding $\left(\widehat{f}_{v_t}(v) - f_{v_t}(v)\right)^2 = o_p\left(n^{-1/2}\right)$, because the convergence rate of the estimator will now only be $\sqrt{T}$. Similarly, a higher order kernel will no longer be needed.

Assumption 31 *$n \to \infty, T \to \infty$, and $T = o(n)$.*

Theorem 2.8.2 *Let Assumption 18, 21, 22, 23, 25, 26, 31, 37, and 39 hold. Assume that bandwidth* $h = c_0 n^{-1/5}$ *in* $\widehat{f}_{v_t}$ *and assume a symmetric kernel of order* $p = 2$*. Then*

$$\sqrt{T}\left[\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{D_{it}Y_{it}}{\widehat{f}_{v_t}(v_{it})}}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{D_{it}}{\widehat{f}_{v_t}(v_{it})}} - \frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{(1-D_{it})Y_{it}}{\widehat{f}_{v_t}(v_{it})}}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{(1-D_{it})}{\widehat{f}_{v_t}(v_{it})}} - E(\widetilde{a}_i + \widetilde{b}_t + Y_1) + E(\widetilde{a}_i + \widetilde{b}_t + Y_0)\right]$$
$$\xrightarrow{d} N\left(0, \frac{var\left(E\left[\Lambda_{1it}|\, b_t, \widetilde{b}_t\right]\right)}{\overline{\Pi}_1^2} - \frac{2cov\left(E\left[\Lambda_{1it}|\, b_t, \widetilde{b}_t\right], E\left[\Lambda_{2it}|\, b_t, \widetilde{b}_t\right]\right)}{\overline{\Pi}_1\overline{\Pi}_2} + \frac{var\left(E\left[\Lambda_{2it}|\, b_t, \widetilde{b}_t\right]\right)}{\overline{\Pi}_2^2}\right).$$

This theorem is proved in the supplemental online appendix.

Remark 2.8.3 In this $\sqrt{T}$ convergence case, we can allow arbitrary dependence between Y_{jit} and $\left(\widetilde{a}_i, \widetilde{b}_t\right)$, which implies that Y_{jit} can contain some general function of $\widetilde{a}_i$ *and* $\widetilde{b}_t$ as long as $E\left(Y_{jit}\right) = E\left(Y_j\right)$, $j = 0, 1$. We similarly allow for more general fixed effects of the form $g\left(\widetilde{a}_i, \widetilde{b}_t\right)$ instead of $\widetilde{a}_i + \widetilde{b}_t$ for some unknown function g, because these fixed effects will still difference out.

Remark 2.8.4 Suppose $\left(a_i, \widetilde{a}_i, b_t, \widetilde{b}_t\right)$ is a series of constants instead of random variables. From the proof of Lemma 2.10.8, the above rate $\sqrt{T}$ limiting distribution will still hold if $\frac{1}{n^2}\left(\sum_{i=1}^{n}\widetilde{a}_i^2\right) = O\left(1\right)$ and $\frac{1}{nT}\left(\sum_{t=1}^{T}\widetilde{b}_t^2\right) = O\left(1\right)$.

2.8.3 Dynamic Panels

Our identification can extend to the dynamic panel case. Define the treatment indicator equation as

$$D_{it} = I(\alpha_0(x_{it}) \leq a_i + b_t + V_{it} + \vartheta\left(D_{it-1}\right) + U_{it} \leq \alpha_1(x_{it})), \tag{2.8.2}$$

and the outcome equation as

$$Y_{it} = \widetilde{a}_i + \widetilde{b}_t + g\left(Y_{it-1}\right) + Y_{0it} + (Y_{1it} - Y_{0it})D_{it}, \tag{2.8.3}$$

where the treatment indicator and the outcome variable are related to those in the last period, and these effects are captured by two unknown functions ϑ, g.

As before, the observables in the model are the outcome Y, treatment D, instrument V, and covariate vector X. $(a_i, b_t, \widetilde{a}_i, \widetilde{b}_t)$ as fixed effects, which can correlate with unobservables and with X in unknown ways.

Assumption 32 *For individuals i and time periods t, $a_i, b_t, \widetilde{a}_i, \widetilde{b}_t$ are random variables.*

$$E\left(g\left(Y_{it-1}\right) + \widetilde{a}_i + \widetilde{b}_t + Y_{jit} | X_{it}, V_{it}, a_i, b_t, U_{it}, D_{it-1}\right) = E\left(g\left(Y_{it-1}\right) + \widetilde{a}_i + \widetilde{b}_t + Y_{jit} \middle| X_{it}, a_i, b_t, U_{it}, D_{it-1}\right), \tag{2.8.4}$$

for $j = 0, 1$, and

$$V_{it} \perp a_i, b_t, U_{it}, D_{it-1} | X_{it}. \tag{2.8.5}$$

Remark 2.8.5 Equation (2.8.4) does not put much more restriction than the corresponding part in Assumption 22. Equation (2.8.5), however, generally requires that $V_{it} \perp V_{it-1}$. Although we impose $V_{it} \perp V_{it-1}$ in Assumption 26, it is for theoretical convenience and could be extended to the weak dependence case. On the contrary, we generally cannot allow any dependence between V_{it} and V_{it-1} for the identification here. If there is no dynamics in the selection equation, i.e., $\vartheta = 0$, then we do not need $V_{it} \perp V_{it-1}$. We would like to emphasize that, just as in previous sections, all strong assumptions are only on the special regressor V.

Assumption 33 *Assumption 20 holds after replacing* $supp[\alpha_0(X) - U, \alpha_1(X) - U]$ *with* $supp[\alpha_0(x_{it}) - \widetilde{a}_i - \widetilde{b}_t - U_{it} - \vartheta(D_{it-1}), \alpha_1(x_{it}) - \widetilde{a}_i - \widetilde{b}_t - U_{it} - \vartheta(D_{it-1})]$.

Theorem 2.8.6 *Let Assumption 18, 32, and 33 hold for each individual* i *in each time period* t. *Let* f_{v_t} *denote the density of* V *in time* t. *Then*

$$\frac{E[D_{it}Y_{it}/f_{v_t}(V_{it}|X_{it})|X_{it}]}{E[D_{it}/f_{v_t}(V_{it}|X_{it})|X_{it}]} - \frac{E[(1-D_{it})Y_{it}/f_{v_t}(V_{it}|X_{it})|X_{it}]}{E[(1-D_{it})/f_{v_t}(V_{it}|X_{it})|X_{it}]} = E(Y_{1it} - Y_{0it}|X_{it}). \quad (2.8.6)$$

It follows the proof as in Theorem 2.3.4.

We provide another set of weaker assumptions that permit some limited dependence among $\{V_{it}\}_{t=1}^T$ (e.g. $\{V_{it}\}_{t=1}^T$ can be a Markov chain, see Remark 2.8.7) and are able to achieve identification. We need to modify the estimator a bit as in equation (2.8.9) where $f_{v_t}(V_{it}|X_{it})$ is replaced by $f_{v_t}(V_{it}|X_{it}, V_{it-1})$.

Assumption 34 *For individuals* i *and time periods* t, $a_i, b_t, \widetilde{a}_i, \widetilde{b}_t$ *are random variables.*

$$\begin{aligned} & E\left(g(Y_{it-1}) + \widetilde{a}_i + \widetilde{b}_t + Y_{jit}|X_{it}, V_{it}, a_i, b_t, U_{it}, D_{it-1}, V_{it-1}\right) \\ = \; & E\left(g(Y_{it-1}) + \widetilde{a}_i + \widetilde{b}_t + Y_{jit}\middle| X_{it}, a_i, b_t, U_{it}, D_{it-1}, V_{it-1}\right), \end{aligned} \quad (2.8.7)$$

for $j = 0, 1$, *and*

$$V_{it} \perp a_i, b_t, U_{it}, D_{it-1} | (X_{it}, V_{it-1}). \quad (2.8.8)$$

Remark 2.8.7 Condition (2.8.8) can be implied by $V_{it} \perp a_i, b_t, \{U_{ij}\}_{j=1}^t \Big| X_{it}$, and $V_{it} \perp \{V_{ij}\}_{j=1}^{t-2} \Big| V_{it-1}$. So $\{V_{it}\}_{t=1}^T$ can be a Markov chain under this condition.

Theorem 2.8.8 *Let Assumption 18, 33 and 34 hold for each individual* i *in each time*

period t. Let f_{v_t} denote the density of V in time t. Then

$$\frac{E[D_{it}Y_{it}/f_{v_t}(V_{it}|X_{it},V_{it-1})|X_{it}]}{E[D_{it}/f_{v_t}(V_{it}|X_{it},V_{it-1})|X_{it}]}-\frac{E[(1-D_{it})Y_{it}/f_{v_t}(V_{it}|X_{it},V_{it-1})|X_{it}]}{E[(1-D_{it})/f_{v_t}(V_{it}|X_{it},V_{it-1})|X_{it}]}=E(Y_{1it}-Y_{0it}|X_{it}). \tag{2.8.9}$$

The proof is in the supplemental Appendix.

Heckman and Navarro (2007) obtain the identification of structural dynamic discrete choice model and models for dynamic treatment effects. Although the model specifications in Heckman and Navarro (2007) and our paper are very different, the identification strategies are similar. Both their paper and ours rely on the sufficient variation on some covariates. Heckman and Navarro (2007) allows more general serial correlation than our paper, because we could only allow limited dependence on the special regressor. However, we model the feedback effects and fixed effects explicitly in the choice and outcome equations, while Heckman and Navarro (2007) does not. Like other sections of our paper, all strong assumptions are on the special regressor.

2.9 Appendix C: Additional Assumptions and Proofs

Proof of Theorem 2.3.2.2 To prove this look first at

$$\begin{aligned} E\left(\frac{I_\tau DY}{f(V \mid X)} \mid U, X\right) &= E\left[E\left(\frac{I_\tau DY_1}{f(V \mid X)} \mid V, U, X\right) \mid U, X\right] \\ &= E\left[\frac{I_\tau I\left[\alpha_0(X) \leq V + U \leq \alpha_1(X)\right] E(Y_1 \mid V, U, X)}{f(V \mid X)} \mid U, X\right] \\ &= \int_{supp(V|U,X)} \frac{I_\tau I\left[\alpha_0(X) - U \leq v \leq \alpha_1(X) - U\right] E(Y_1 \mid U, X)}{f(v \mid X)} f(v \mid U, X)\, dv \\ &= \int_{\alpha_0(X)-U}^{\alpha_1(X)-U} \frac{E(Y_1 \mid U, X)}{f(v \mid X)} f(v \mid X)\, dv = E(Y_1 \mid U, X) \int_{\alpha_0(X)-U}^{\alpha_1(X)-U} 1 dv \\ &= \left[\alpha_1(X) - \alpha_0(X)\right] E(Y_1 \mid U, X), \end{aligned}$$

the fourth equality holds by Assumption 20.

Therefore

$$E\left(\frac{I_\tau DY}{f(V \mid X)} \mid X\right) = \left[\alpha_1(X) - \alpha_0(X)\right] E(Y_1 \mid X)$$

The same analysis dropping Y gives

$$E\left(\frac{I_\tau D}{f(V \mid X)} \mid X\right) = \alpha_1(X) - \alpha_0(X)$$

so

$$E\left(\frac{I_\tau DY}{f(V \mid X)} \mid X\right) = E(Y_1 \mid X)\, E\left(\frac{I_\tau D}{f(V \mid X)} \mid X\right)$$

Similarly,

$$\begin{aligned}
E\left(\frac{I_\tau\left(1-D\right)Y}{f\left(V\mid X\right)}\mid X\right) &= E\left(\frac{I_\tau\left(1-D\right)Y_0}{f\left(V\mid X\right)}\mid X\right)\\
&= E\left(\frac{I_\tau Y_0}{f\left(V\mid X\right)}\mid X\right)-E\left(\frac{I_\tau DY_0}{f\left(V\mid X\right)}\mid X\right)\\
&= E\left(Y_0\mid X\right)E\left(\frac{I_\tau}{f\left(V\mid X\right)}\mid X\right)-\left[\alpha_1\left(X\right)-\alpha_0\left(X\right)\right]E\left(Y_0\mid X\right)\\
&= E\left(Y_0\mid X\right)E\left(\frac{I_\tau}{f\left(V\mid X\right)}-\left[\alpha_1\left(X\right)-\alpha_0\left(X\right)\right]\mid X\right)\\
&= E\left(Y_0\mid X\right)E\left(\frac{I_\tau\left(1-D\right)}{f\left(V\mid X\right)}\mid X\right)
\end{aligned}$$

Together these equations prove the result. ■

Proof of Theorem 2.3.4.2 The proof is the almost the same as the proof for Theorem 2.3.2. To prove this first look at

$$\begin{aligned}
&E\left(\frac{I_{\tau it}D_{it}Y_{it}}{f_{v_t}\left(V_{it}|X_{it}\right)}\middle|U_{it},a_i,b_t,X_{it}\right)\\
=\;& E\left[E\left(\frac{I_{\tau it}D_{it}\left(\widetilde{a}_i+\widetilde{b}_t+Y_{1it}\right)}{f_{v_t}\left(V_{it}|X_{it}\right)}\mid V_{it},U_{it},a_i,b_t,X_{it}\right)\mid U_{it},a_i,b_t,X_{it}\right]\\
=\;& E\left[\frac{I_{\tau it}I\left(\alpha_0(X_{it})\leq a_i+b_t+V_{it}+U_{it}\leq\alpha_1(X_{it})\right)E\left(\widetilde{a}_i+\widetilde{b}_t+Y_{1it}\mid V_{it},U_{it},a_i,b_t,X_{it}\right)}{f_{v_t}\left(V_{it}|X_{it}\right)}\mid U_{it},a_i,b_t,X_{it}\right]\\
=\;& \int_{supp(V_{it}|U_{it},a_i,b_t,X_{it})}\frac{I_{\tau it}I\left(\alpha_0(X_{it})-a_i-b_t-U_{it}\leq v_{it}\leq\alpha_1(X_{it})-a_i-b_t-U_{it}\right)}{f_{v_t}\left(v_{it}|X_{it}\right)}\\
&E\left(\widetilde{a}_i+\widetilde{b}_t+Y_{1it}\mid U_{it},a_i,b_t,X_{it}\right)f_{v_t}\left(v_{it}\mid U_{it},a_i,b_t,X_{it}\right)dv_{it}\\
=\;& \int_{\alpha_0(X_{it})-a_i-b_t-U_{it}}^{\alpha_1(X_{it})-a_i-b_t-U_{it}}\frac{E\left(\widetilde{a}_i+\widetilde{b}_t+Y_{1it}\mid U_{it},a_i,b_t,X_{it}\right)}{f_{v_t}\left(v_{it}|X_{it}\right)}f_{v_t}\left(v_{it}|X_{it}\right)dv_{it}\\
=\;& E\left(\widetilde{a}_i+\widetilde{b}_t+Y_{1it}\mid U_{it},a_i,b_t,X_{it}\right)\int_{\alpha_0(X_{it})-a_i-b_t-U_{it}}^{\alpha_1(X_{it})-a_i-b_t-U_{it}}1dv_{it}\\
=\;& E\left(\widetilde{a}_i+\widetilde{b}_t+Y_{1it}\mid U_{it},a_i,b_t,X_{it}\right)\left[\alpha_1(X_{it})-\alpha_0(X_{it})\right]
\end{aligned}$$

and therefore

$$\begin{aligned} & E\left[I_{\tau it} D_{it} Y_{it} / f_{v_t}\left(V_{it} | X_{it}\right) | X_{it}\right] \\ = \; & E\left[E\left(\widetilde{a}_i + \widetilde{b}_t + Y_{1it} \mid U_{it}, a_i, b_t, X_{it}\right)\left[\alpha_1(X_{it}) - \alpha_0(X_{it})\right] | X_{it}\right] \\ = \; & E\left(Y_{1it} + \widetilde{a}_i + \widetilde{b}_t \middle| X_{it}\right)\left[\alpha_1(X_{it}) - \alpha_0(X_{it})\right]. \end{aligned}$$

Given the above result, the rest of the proof follows similarly as in the proof for Theorem 2.3.2. ■

We let $\mathbf{m}_k \equiv (m_1, m_2, ..., m_k)$ be a $k \times 1$ non-negative integers. Following Masry (1996), we adopt the notation: $u^{\mathbf{m}_k} \equiv \Pi_{i=1}^k u_i^{m_i}$, $\mathbf{m}_k! \equiv \Pi_{i=1} m_i!$, $|\mathbf{m}_k| \equiv \sum_{i=1}^k m_i$, and $\sum_{|\mathbf{m}_k|=p} \equiv \sum_{\substack{m_1=0 \\ m_1+\cdots+m_k=p}}^{p} \cdots \sum_{m_k=0}^{p}$. We let $D^{\mathbf{m}_k} f_x(x) \equiv \partial^{|\mathbf{m}_k|} f_x(x) / \partial^{m_1} x_1 \cdots \partial^{m_k} x_k$. If we have covariates of other dimensions, e.g. $k+1$, then $\mathbf{m}_{k+1}$ and the other notations above are changed accordingly with k replaced by $k+1$.

Assumption 35 *Observations are i.i.d. across i.*

Assumption 36 *$f_x(x)$, $E(g_{1i}|x)$, and $E(g_{2i}|x)$ are bounded away from zero over the whole support of X.*

Assumption 37 *The kernel functions $K(v)$, $K(x)$, and $K(x,v)$ have supports that are convex and bounded on $\mathbb{R}^1$, $\mathbb{R}^k$, and $\mathbb{R}^{k+1}$ respectively. Each kernel function integrates to one over its support, is symmetric around zero, and has order p, i.e., for $K(x)$,*

$$\int_{\mathbb{R}^k} x^{\mathbf{m}_k} K(x) dx = 0 \quad for \;\; |\mathbf{m}_k| < p,$$

$$\int_{\mathbb{R}^k} x^{\mathbf{m}_k} K(x) dx \neq 0 \;\; for \; some \;\; |\mathbf{m}_k| = p,$$

and $\int K(x)^2 dx, \int_{\mathbb{R}^k} |x^{\mathbf{m}_k}| K(x) dx$ *for* $|\mathbf{m}_k| = p$ *are finite. This similarly holds for* $K(v)$ *and* $K(x, v)$.

Assumption 38 *Let* $s_{1i} \equiv \frac{D_i I_{\tau i} Y_i}{f_{xv}(x_i, v_i)}$, $s_{2i} \equiv \frac{D_i I_{\tau i} Y_i f_x(x_i)}{f_{xv}^2(x_i, v_i)}$, $s_{3i} \equiv \frac{D_i I_{\tau i}}{f_{xv}(x_i, v_i)}$, $s_{4i} \equiv \frac{D_i I_{\tau i} f_x(x_i)}{f_{xv}^2(x_i, v_i)}$, $s_{5i} \equiv \frac{(1-D_i) I_{\tau i} Y_i}{f_{xv}(x_i, v_i)}$, $s_{6i} \equiv \frac{(1-D_i) I_{\tau i} Y_i f_x(x_i)}{f_{xv}^2(x_i, v_i)}$, $s_{7i} \equiv \frac{(1-D_i) I_{\tau i}}{f_{xv}(x_i, v_i)}$, $s_{8i} \equiv \frac{(1-D_i) I_{\tau i} f_x(x_i)}{f_{xv}^2(x_i, v_i)}$. *Then for each* s_{ji}, $j = 1, ..., 8$, $f_x(x)$, $f_{xv}(x, v)$ *satisfy the Lipschitz condition that there exists some positive numbers* $M_1, ...M_{10}$, *such that*

$$|E(s_{ji}|x + e_x) - E(s_{ji}|x)| \leq M_j \|e_x\|, \ j = 1, ..., 8$$

$$\begin{aligned} |f_x(x + e_x) - f_x(x)| &\leq M_9 \|e_x\|, \\ |f_{xv}(x + e_x, v + e_v) - f_x(x, v)| &\leq M_{10} \|(e_x, e_v)\|. \end{aligned}$$

$E(s_{ji}|x_i)$, $j = 1, ..., 8$, f_x, f_{xv} *are p-th order differentiable and the p-th order derivatives are bounded. The p-th order derivatives of* f_x, f_{xv} *also satisfy the Lipschitz condition. The second moment of* $q_i(x)$ *(defined in equation 2.10.6) exists.*

Assumption 39 $E(D_{it} I_{\tau it} Y_{it}|v)$, $E[(1 - D_{it}) I_{\tau it} Y_{it}|v]$, $f_{v_t}(v)$ *are p times continuous differentiable in* v, *and the p-th order derivatives are bounded. Second moments of* $\frac{D_{it} I_{\tau it} Y_{it}}{f_{v_t}(v_{it})}$, $\frac{D_{it} I_{\tau it}}{f_{v_t}(v_{it})}$, $\frac{(1-D_{it}) I_{\tau it} Y_{it}}{f_{v_t}(v_{it})}$, *and* $\frac{(1-D_{it}) I_{\tau it} Y_{it}}{f_{v_t}(v_{it})}$ *are bounded.*

Table 1: Summary Statistics of the US Dataset

	MEAN	SD	LQ	MED	UQ
Competition	0.76	0.11	0.70	0.76	0.83
Innovation	5.53	9.98	0.22	1.59	5.77
Source-weighted Interest Rate	0.91	0.23	0.79	0.87	0.99

Note: MEAN = mean. SD = standard errors. LQ = 25% quantile (lower). MED = 50% quantile (median). UQ = 75% quantile (upper).

Table 2A: Empirical Estimates in Various Cases

	Right Threshold	Left Threshold	Trim-ATE	No-Trim-ATE	Naive-ATE	ML-ATE
Case 1	25%(0.70)	75%(0.83)	-3.90 (0.61)	-4.25 (0.75)	-1.89 (0.27)	-1.85 (0.39)
Case 2	33%(0.72)	67%(0.80)	-3.27 (0.52)	-3.47 (0.66)	-1.67 (0.26)	-1.69 (0.37)
Case 3	10%(0.63)	90%(0.89)	-2.77 (0.98)	-2.75 (1.10)	-1.95 (0.29)	-4.40 (3.48)
Case 4	20%(0.68)	80%(0.85)	-4.25 (0.71)	-4.62 (0.86)	-2.22 (0.28)	-2.12 (0.43)
Case 5	30%(0.71)	70%(0.82)	-3.54 (0.54)	-3.95 (0.68)	-1.83 (0.26)	-1.81 (0.37)
Case 6	40%(0.74)	60%(0.79)	-2.49 (0.54)	-2.58 (0.67)	-1.18 (0.25)	-1.48 (0.39)

Notes: Right Threshold and Left Threshold are the $\underline{\alpha}$ and $\overline{\alpha}$ in Equation (2.4.5) respectively. The first value is the percentage of competition set for the thresholds, with corresponding value of competition in the parenthesis. Four different estimates are reported here, with standard errors in parenthesis. Trim-ATE and No-Trim-ATE are our proposed estimator with and without trimming (2%) respectively. Naive-ATE is an estimate for $E(Y_1|T=1) - E(Y_0|T=0)$. ML-ATE is Heckman's selection MLE.

Table 2B: Empirical Estimates in Various Cases - Robust Check

	Right Threshold	Left Threshold	Trim-ATE	No-Trim-ATE	Naive-ATE	ML-ATE
Case 1	25%	75%	-4.02 (0.63)	-4.29 (0.80)	-2.04 (0.27)	-2.02 (0.39)
Case 2	33%	67%	-3.46 (0.53)	-4.02 (0.66)	-1.81 (0.26)	-4.46 (0.64)
Case 3	10%	90%	-3.05 (1.06)	-2.98 (1.20)	-2.26 (0.29)	-4.51 (3.00)
Case 4	20%	80%	-4.98 (0.74)	-5.03 (0.93)	-2.75 (0.28)	-2.69 (0.44)
Case 5	30%	70%	-3.62 (0.56)	-3.86 (0.70)	-1.86 (0.26)	-5.95 (0.50)
Case 6	40%	60%	-2.41 (0.57)	-2.99 (0.67)	-0.99 (0.26)	-0.97 (0.44)

Notes: Right Threshold and Left Threshold are the $\underline{\alpha}$ and $\overline{\alpha}$ in Equation (2.4.5) respectively. Four different estimates are reported here, with standard errors in parenthesis. Trim-ATE and No-Trim-ATE are our proposed estimator with and without trimming (2%) respectively. Naive-ATE is an estimate for $E(Y_1|T=1) - E(Y_0|T=0)$. ML-ATE is Heckman's selection MLE.

Table 3A: P-Value of the Testing of the Large Support Assumption

Year	Left	Right	Year	Left	Right
1976	0.000***	0.000***	1977	0.000***	0.001***
1978	1.000	0.000***	1979	0.000***	0.000***
1980	0.051*	0.000***	1981	1.000	0.000***
1982	1.000	1.000	1983	0.592	0.000***
1984	0.008***	0.000***	1985	0.000***	0.000***
1986	0.000***	0.000***	1987	0.658	0.010***
1988	0.846	0.000***	1989	0.000***	0.000***
1990	0.000***	0.000***	1991	0.000***	0.000***
1992	0.000***	0.000***	1993	1.000	0.000***
1994	0.000***	1.000	1995	0.000***	1.000
1996	0.867	1.000	1997	0.000***	0.608
1998	0.000***	0.000***	1999	0.000***	0.725
2000	0.000***	0.347	2001	0.000***	0.000***

Notes: ***, **, * denote the cases when P-values are less than 0.01, 0.05, 0.10 respectively.

Table 3B: Ordered Choice Estimates (Identification at Infinity)

Trimming Parameter	50%	40%	30%	20%	10%	5%
	Panel A: Define D from the whole sample					
$E(Y_0)$	10.17 (0.97)	11.35 (1.17)	12.78 (1.46)	15.56 (1.98)	17.40 (2.87)	26.94 (4.44)
$E(Y_2)$	5.86 (0.50)	5.96 (0.58)	5.94 (0.67)	5.94 (0.88)	4.56 (0.76)	5.06 (0.98)
	Panel B: Define D separately each year (robust check)					
$E(Y_0)$	9.61 (0.95)	10.55 (1.17)	11.67 (1.41)	15.05 (2.08)	14.85 (2.42)	18.90 (4.53)
$E(Y_2)$	6.55 (0.49)	6.39 (0.53)	6.54 (0.62)	6.07 (0.77)	5.79 (0.95)	7.96 (1.76)

Notes: The estimates are obtained from the identification at infinity, with standard deviation in parentheses.
The choice of the trimming parameters is based on the specified percentages of uncensored observations.

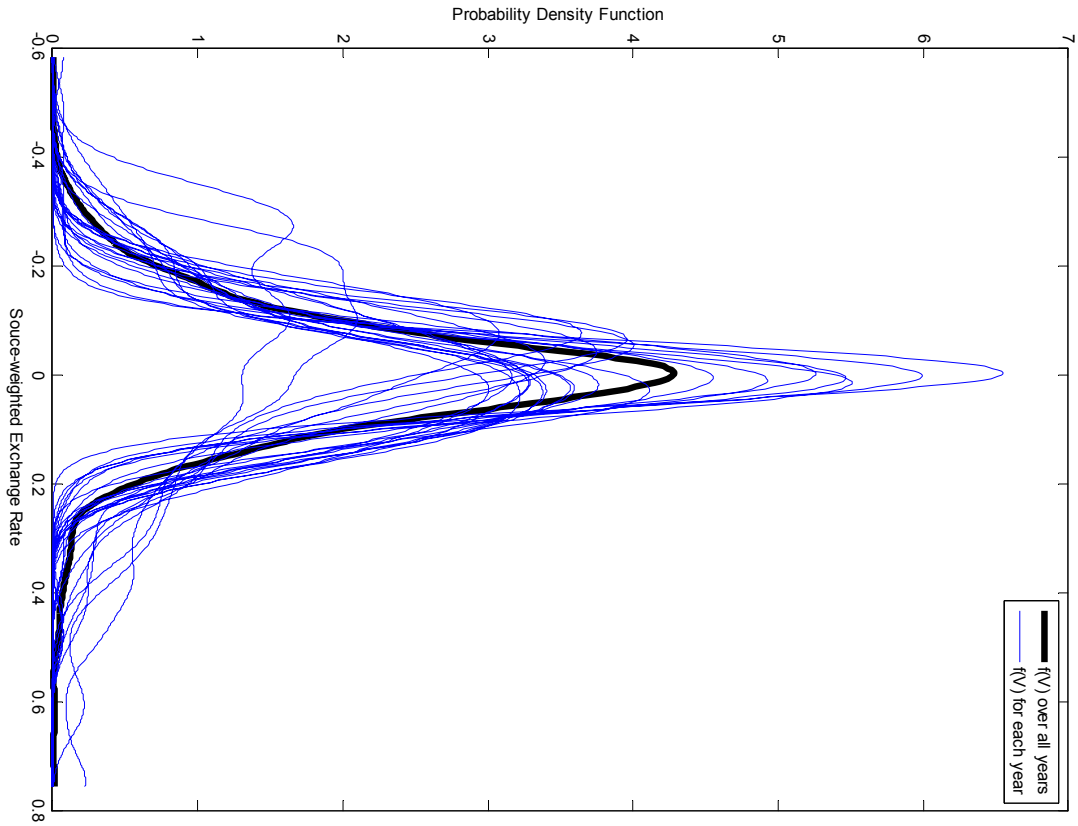

Figure 1: Distribution of V

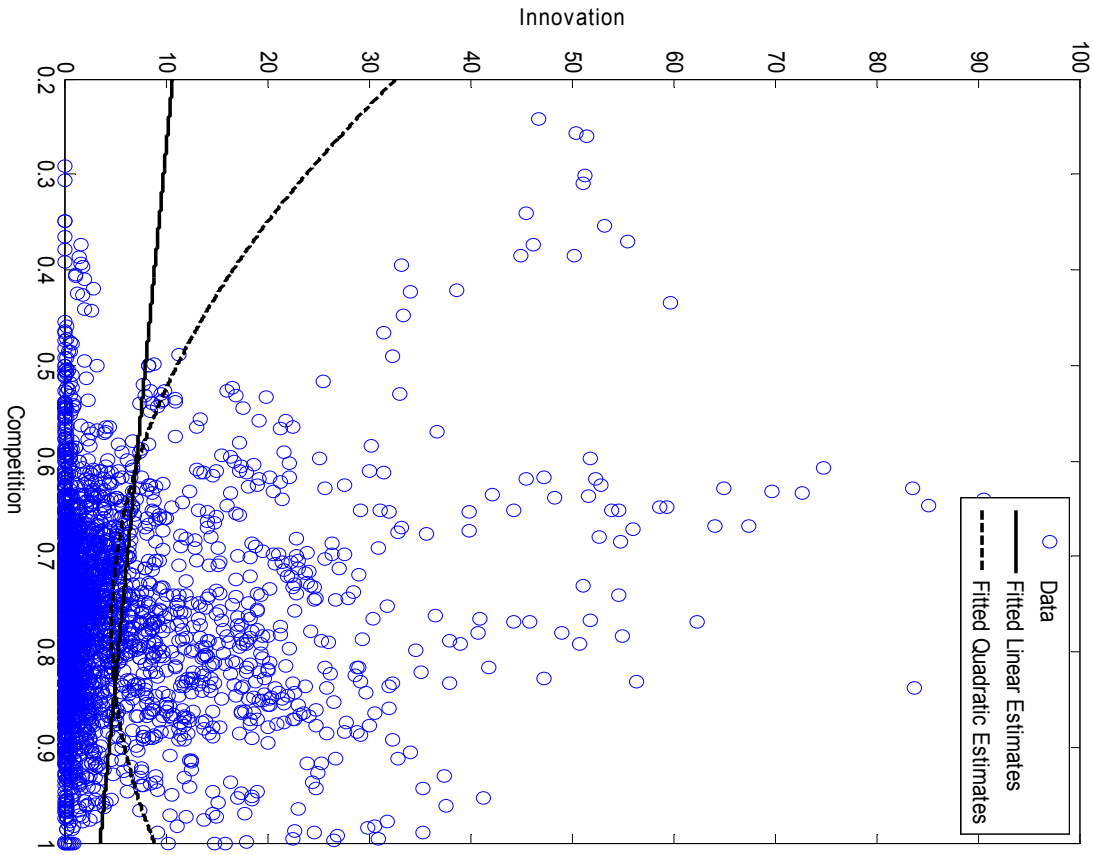

Figure 2: Competition and Innovation

2.10 Appendix D: Supplemental Appendix

This supplemental appendix provides proofs for Theorem 2.3.5, 2.3.9, 2.8.2 in Section 2.10.1, proofs for Theorem 2.5.1, 2.5.3, 2.5.5 in Section 2.10.2, proofs for Theorem 2.8.1, 2.8.8 in Section 2.10.3, and Table 4, 5, and 6 in Section 2.10.4.

To make the proof more clearly, we suppress the trimming indicators $I_{\tau i}, I_{\tau it}$ in the supplemental appendix. The proof can still go through when the trimming indicators are present.

Remark 2.10.1 (Uniform Convergence) Based on Silverman (1978), we have the uniform convergence of $\widehat{f}_{xv}(x,v)$ and $\widehat{f}_{x}(x)$ over a compact set of (V,X) and X respectively. We could use this result for those observations in our estimator (2.3.5) with nonzero weight for the following reasons. For the estimation at x, an interior point of the support of X, because we use a kernel function K with bounded support, those x_i outside of a small interval around x will have zero weights. When h is small enough, those x_i with non-zero weights will eventually fall into the compact set where we have the uniform convergence results. For the estimation at v, we put a trimming indicator as in the identification theorem. If we select a compact set that strictly covers the one where that trimming indicator is nonzero, we have the uniform convergence results for all v_i with nonzero weights.

2.10.1 Proof of Theorem 2.3.5 and 2.3.9, 2.8.2

We let $\widehat{h}_{1i} = \frac{D_i Y_i}{\widehat{f}(v_i|x_i)}$, $\widehat{g}_{1i} = \frac{D_i}{\widehat{f}(v_i|x_i)}$,where $\widehat{f}(v_i|x_i) = \frac{\widehat{f}_{xv}(x_i,v_i)}{\widehat{f}_x(x_i)}$, and $\widehat{f}_x(x_i)$ and $\widehat{f}_{xv}(x_i,v_i)$ are standard leave-one-out nonparametric density estimators:

$$\begin{aligned}
\widehat{f}_x(x_i) &= \frac{1}{nh^k}\sum_{l=1,l\neq i}^{n} K\left(\frac{x_l - x_i}{h}\right), \\
\widehat{f}_{xv}(x_i,v_i) &= \frac{1}{nh^{k+1}}\sum_{l=1,l\neq i}^{n} K\left(\frac{x_l - x_i}{h}, \frac{v_l - v_i}{h}\right).
\end{aligned}$$

where h is the bandwidth and K is the kernel function. Without loss of generality, we use the same h for each covariate. The kernel function K is defined in Assumption 37.

The sample counterpart estimate for $\psi_1(x)$ could be then

$$\widehat{\psi}_1(x) = \frac{\widehat{E}\left(\widehat{h}_{1i}\middle| x\right)}{\widehat{E}\left(\widehat{g}_{1i}\middle| x\right)}, \tag{2.10.1}$$

where $\widehat{E}$ denotes the standard kernel nonparametric estimation:

$$\begin{aligned}
\widehat{E}\left(\widehat{h}_{1i}\middle| x\right) &= \frac{1}{nh^k}\sum_{i=1}^{n}\widehat{h}_{1i}K\left(\frac{x_i - x}{h}\right)\Bigg/\left[\frac{1}{nh^k}\sum_{i=1}^{n}K\left(\frac{x_i - x}{h}\right)\right], \\
\widehat{E}\left(\widehat{g}_{1i}\middle| x\right) &= \frac{1}{nh^k}\sum_{i=1}^{n}\widehat{g}_{1i}K\left(\frac{x_i - x}{h}\right)\Bigg/\left[\frac{1}{nh^k}\sum_{i=1}^{n}K\left(\frac{x_i - x}{h}\right)\right].
\end{aligned}$$

For simplicity, we abuse the notation a bit by defining

$$\widetilde{h}_{1i} \equiv h_{1i} f_x(x_i),\ \widetilde{g}_{1i} \equiv g_{1i} f_x(x_i) \tag{2.10.2}$$

and $\widehat{E}\left(\widehat{\widetilde{h}}_{1i}\middle| x\right)$ and $\widehat{E}\left(\widehat{\widetilde{g}}_{1i}\middle| x\right)$ are defined as the numerators in $\widehat{E}\left(\widehat{h}_{1i}\middle| x\right)$ and $\widehat{E}\left(\widehat{g}_{1i}\middle| x\right)$

respectively:

$$\widehat{E}\left(\widehat{\widetilde{h}}_{1i}\middle|\, x\right) \equiv \frac{1}{nh^k}\sum_{i=1}^{n}\widehat{h}_{1i}K\left(\frac{x_i-x}{h}\right), \tag{2.10.3}$$

$$\widehat{E}\left(\widehat{\widetilde{g}}_{1i}\middle|\, x\right) \equiv \frac{1}{nh^k}\sum_{i=1}^{n}\widehat{g}_{1i}K\left(\frac{x_i-x}{h}\right). \tag{2.10.4}$$

It follows from the definition of $\widetilde{h}_{1i}$ and $\widetilde{g}_{1i}$ that

$$E\left(\widetilde{h}_{1i}\middle|\, x\right) = E\left(h_{1i}|\, x\right) f_x\left(x\right) \text{ and } E\left(\widetilde{g}_{1i}|\, x\right) = E\left(g_{1i}|\, x\right) f_x\left(x\right),$$

and $\widehat{\psi}_1(x) = \frac{\widehat{E}\left(\widehat{\widetilde{h}}_{1i}\middle| x\right)}{\widehat{E}\left(\widehat{\widetilde{g}}_{1i}\middle| x\right)}$.

Replacing the subscript 1 with 2, similarly define $\widehat{\psi}_2(x)$, $\widehat{E}\left(\widehat{h}_{2i}\middle|\, x\right)$, $\widehat{E}\left(\widehat{g}_{2i}|\, x\right)$, $\widetilde{h}_{2i}$, $\widetilde{g}_{2i}$, $\widehat{\widetilde{h}}_{2i}$, $\widehat{\widetilde{g}}_{2i}$, $\widehat{E}\left(\widehat{\widetilde{h}}_{2i}\middle|\, x_i\right)$, $\widehat{E}\left(\widehat{\widetilde{g}}_{2i}\middle|\, x_i\right)$. The resulting estimator is then

$$\begin{aligned}
\widehat{\psi}_1(x) - \widehat{\psi}_2(x) &= \frac{\frac{1}{nh^k}\sum_{i=1}^{n}\frac{D_iY_i}{\widehat{f}(v_i|x_i)}K\left(\frac{x_i-x}{h}\right)}{\frac{1}{nh^k}\sum_{i=1}^{n}\frac{D_i}{\widehat{f}(v_i|x_i)}K\left(\frac{x_i-x}{h}\right)} - \frac{\frac{1}{nh^k}\sum_{i=1}^{n}\frac{(1-D_i)Y_i}{\widehat{f}(v_i|x_i)}K\left(\frac{x_i-x}{h}\right)}{\frac{1}{nh^k}\sum_{i=1}^{n}\frac{1-D_i}{\widehat{f}(v_i|x_i)}K\left(\frac{x_i-x}{h}\right)} \\
&= \frac{\widehat{E}\left(\widehat{h}_{1i}\middle|\, x\right)}{\widehat{E}\left(\widehat{g}_{1i}|\, x\right)} - \frac{\widehat{E}\left(\widehat{h}_{2i}\middle|\, x\right)}{\widehat{E}\left(\widehat{g}_{2i}|\, x\right)} = \frac{\widehat{E}\left(\widehat{\widetilde{h}}_{1i}\middle|\, x\right)}{\widehat{E}\left(\widehat{\widetilde{g}}_{1i}\middle|\, x\right)} - \frac{\widehat{E}\left(\widehat{\widetilde{h}}_{2i}\middle|\, x\right)}{\widehat{E}\left(\widehat{\widetilde{g}}_{2i}\middle|\, x\right)}.
\end{aligned} \tag{2.10.5}$$

We define the following term for the influence function of $\widehat{\psi}_1(x_i) - \widehat{\psi}_2(x_i)$:

$$
\begin{aligned}
q_i(x) \equiv & \left(\frac{h_{1i}}{E(\widetilde{g}_{1i}|x)} + \frac{E(h_{1i}|x_i)}{E(\widetilde{g}_{1i}|x)} - \frac{E(h_{1i}|x_i, v_i)}{E(\widetilde{g}_{1i}|x)} - \frac{E\left(\widetilde{h}_{1i}\middle| x\right) g_{1i}}{E(\widetilde{g}_{1i}|x)^2} - \frac{E\left(\widetilde{h}_{1i}\middle| x\right) E(g_{1i}|x_i)}{E(\widetilde{g}_{1i}|x)^2} \right. \\
& \left. + \frac{E\left(\widetilde{h}_{1i}\middle| x\right) E(g_{1i}|x_i, v_i)}{E(\widetilde{g}_{1i}|x)^2} \right) - \left(\frac{h_{2i}}{E(\widetilde{g}_{2i}|x)} + \frac{E(h_{2i}|x_i)}{E(\widetilde{g}_{2i}|x)} - \frac{E(h_{2i}|x_i, v_i)}{E(\widetilde{g}_{2i}|x)} - \frac{E\left(\widetilde{h}_{2i}\middle| x\right) g_{2i}}{E(\widetilde{g}_{2i}|x)^2} \right. \\
& \left. - \frac{E\left(\widetilde{h}_{2i}\middle| x\right) E(g_{2i}|x_i)}{E(\widetilde{g}_{2i}|x)^2} + \frac{E\left(\widetilde{h}_{2i}\middle| x\right) E(g_{2i}|x_i, v_i)}{E(\widetilde{g}_{2i}|x)^2} \right). \qquad (2.10.6)
\end{aligned}
$$

The bias term resulted from nonparametric regression is defined by:

$$
\begin{aligned}
\mathbb{B}_p(x) \equiv & \frac{\mathbb{B}_{1,p}}{E(g_{1i}|x)} - \frac{\mathbb{B}_{2,p}}{E(g_{1i}|x)} - \frac{E(h_{1i}|x)\,\mathbb{B}_{3,p}}{E(g_{1i}|x)^2} + \frac{E(h_{1i}|x)\,\mathbb{B}_{4,p}}{E(g_{1i}|x)^2} \qquad (2.10.7) \\
& - \frac{\mathbb{B}_{5,p}}{E(g_{2i}|x)} + \frac{\mathbb{B}_{6,p}}{E(g_{2i}|x)} + \frac{E(h_{2i}|x)\,\mathbb{B}_{7,p}}{E(g_{2i}|x)^2} - \frac{E(h_{2i}|x)\,\mathbb{B}_{8,p}}{E(g_{2i}|x)^2},
\end{aligned}
$$

where $\mathbb{B}_{j,p}$, $j = 1, ..., 8$, are defined in equation (2.10.8).

$$
\begin{aligned}
\mathbb{B}_{1,p} &\equiv h^p \sum_{|\mathbf{m}_k|=p} \frac{E\left[D_i Y_i / f_{xv}\left(x_i, v_i\right) D^{\mathbf{m}_k} f_x\left(x_i\right) \middle| x\right]}{\mathbf{m}_k!} \int_{\mathbb{R}^k} u_l^{\mathbf{m}_k} K\left(u_l\right) du_l, \qquad (2.10.8)\\
\mathbb{B}_{2,p} &\equiv h^p \sum_{|\mathbf{m}_{k+1}|=p} \frac{E\left[D_i Y_i f_x\left(x_i\right) / f_{xv}^2\left(x_i, v_i\right) D^{\mathbf{m}_{k+1}} f_{xv}\left(x_i, v_i\right) \middle| x\right]}{\mathbf{m}_{k+1}!} \int_{\mathbb{R}^{k+1}} u_l^{\mathbf{m}_{k+1}} K\left(u_l\right) du_l,\\
\mathbb{B}_{3,p} &\equiv h^p \sum_{|\mathbf{m}_k|=p} \frac{E\left[D_i / f_{xv}\left(x_i, v_i\right) D^{\mathbf{m}_k} f_x\left(x_i\right) \middle| x\right]}{\mathbf{m}_k!} \int_{\mathbb{R}^k} u_l^{\mathbf{m}_k} K\left(u_l\right) du_l,\\
\mathbb{B}_{4,p} &\equiv h^p \sum_{|\mathbf{m}_{k+1}|=p} \frac{E\left[D_i f_x\left(x_i\right) / f_{xv}^2\left(x_i, v_i\right) D^{\mathbf{m}_{k+1}} f_{xv}\left(x_i, v_i\right) \middle| x\right]}{\mathbf{m}_{k+1}!} \int_{\mathbb{R}^{k+1}} u_l^{\mathbf{m}_{k+1}} K\left(u_l\right) du_l,\\
\mathbb{B}_{5,p} &\equiv h^p \sum_{|\mathbf{m}_k|=p} \frac{E\left[\left(1 - D_i\right) Y_i / f_{xv}\left(x_i, v_i\right) D^{\mathbf{m}_k} f_x\left(x_i\right) \middle| x\right]}{\mathbf{m}_k!} \int_{\mathbb{R}^k} u_l^{\mathbf{m}_k} K\left(u_l\right) du_l,\\
\mathbb{B}_{6,p} &\equiv h^p \sum_{|\mathbf{m}_{k+1}|=p} \frac{E\left[\left(1 - D_i\right) Y_i f_x\left(x_i\right) / f_{xv}^2\left(x_i, v_i\right) D^{\mathbf{m}_{k+1}} f_{xv}\left(x_i, v_i\right) \middle| x\right]}{\mathbf{m}_{k+1}!} \int_{\mathbb{R}^{k+1}} u_l^{\mathbf{m}_{k+1}} K\left(u_l\right) du_l,\\
\mathbb{B}_{7,p} &\equiv h^p \sum_{|\mathbf{m}_k|=p} \frac{E\left[\left(1 - D_i\right) / f_{xv}\left(x_i, v_i\right) D^{\mathbf{m}_k} f_x\left(x_i\right) \middle| x\right]}{\mathbf{m}_k!} \int_{\mathbb{R}^k} u_l^{\mathbf{m}_k} K\left(u_l\right) du_l, ,\\
\mathbb{B}_{8,p} &\equiv h^p \sum_{|\mathbf{m}_{k+1}|=p} \frac{E\left[\left(1 - D_i\right) f_x\left(x_i\right) / f_{xv}^2\left(x_i, v_i\right) D^{\mathbf{m}_{k+1}} f_{xv}\left(x_i, v_i\right) \middle| x\right]}{\mathbf{m}_{k+1}!} \int_{\mathbb{R}^{k+1}} u_l^{\mathbf{m}_{k+1}} K\left(u_l\right) du_l,
\end{aligned}
$$

Lemma 2.10.2 *Assume we observe* $\underset{(k+1)\times 1}{W_i} = (\underset{k\times 1}{X_i}\ \underset{1\times 1}{V_i})$, $\underset{1\times 1}{s_i}$, $\underset{(k+2)\times 1}{Z_i} = (\underset{(k+1)\times 1}{W_i}\ \underset{1\times 1}{s_i})$, , *which are i.i.d. across i. The density functions* f_x, f_w *for* X *and* W *are bounded.* f_x *and* f_w *are p-th order differentiable, and p-th order derivatives are bounded.* $E(s_i|w_i)$, f_x, f_w *satisfy the Lipschitz condition*

$$|E(s_i|w_i + e_w) - E(s_i|w_i)| \leq M_1 \|e_w\|,$$

$$|f_x(x_i + e_x) - f_x(x_i)| \leq M_2 \|e_x\|,$$

$$|f_w(w_i + e_w) - f_w(w_i)| \leq M_3 \|e_w\|$$

for some positive M_1, M_2, M_3. *Under the above assumptions, when* x *is a interior point of*

X, *we have*

$$\begin{aligned}
&\widehat{E}\left(s_i\widehat{f}_w(w_i)\Big|\, x\right)\widehat{f}_x(x) \qquad (2.10.9)\\
=\;& \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^k}s_i\left[\frac{1}{n-1}\sum_{l=1,l\neq i}^{n}\frac{1}{h^{k+1}}K\left(\frac{w_l-w_i}{h}\right)\right]K\left(\frac{x_i-x}{h}\right)\\
=\;& \frac{1}{n}\sum_{i=1}^{n}\left[\frac{s_i}{h^k}K\left(\frac{x_i-x}{h}\right)f_w\left(w_i\right)+\frac{E\left(s_i|w_i\right)}{h^k}K\left(\frac{x_i-x}{h}\right)f_w\left(w_i\right)\right.\\
&\left.-2E\left(\frac{s_i}{h^k}K\left(\frac{x_i-x}{h}\right)f_w\left(w_i\right)\right)\right]+E\left[\frac{s_i}{h^{2k+1}}K\left(\frac{x_i-x}{h}\right)K(\frac{w_l-w_i}{h})\right]+o_P\left(\frac{1}{\sqrt{nh^k}}\right) \qquad (2.10.10)
\end{aligned}$$

and

$$\begin{aligned}
&\widehat{E}\left(s_i\widehat{f}_x(x_i)\Big|\, x\right)\widehat{f}_x(x) \qquad (2.10.11)\\
=\;& \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^k}s_i\left[\frac{1}{n-1}\sum_{l=1,l\neq i}^{n}\frac{1}{h^{k+1}}K\left(\frac{x_l-x_i}{h}\right)\right]K\left(\frac{x_i-x}{h}\right)\\
=\;& \frac{1}{n}\sum_{i=1}^{n}\left[\frac{s_i}{h^k}K\left(\frac{x_i-x}{h}\right)f_x\left(x_i\right)+\frac{E\left(s_i|x_i\right)}{h^k}K\left(\frac{x_i-x}{h}\right)f_x\left(x_i\right)\right.\\
&\left.-2E\left(\frac{s_i}{h^k}K\left(\frac{x_i-x}{h}\right)f_x\left(x_i\right)\right)\right]+E\left[\frac{s_i}{h^{2k}}K\left(\frac{x_i-x}{h}\right)K(\frac{x_l-x_i}{h})\right]+o_P\left(\frac{1}{\sqrt{nh^k}}\right) \qquad (2.10.12)
\end{aligned}$$

Proof of Lemma 2.10.2.2 Consider first the following term,

$$\begin{aligned}
&\frac{1}{n(n-1)h^{2k+1}}\sum_{i=1}^{n}\sum_{l=1,l\neq i}^{n}s_iK\left(\frac{x_i-x}{h}\right)K\left(\frac{w_l-w_i}{h}\right) \qquad (2.10.13)\\
=\;& \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{l=i+1}^{n}\frac{1}{2}\left[s_iK\left(\frac{x_i-x}{h}\right)+s_lK\left(\frac{x_l-x}{h}\right)\right]\frac{1}{h^{2k+1}}K\left(\frac{w_l-w_i}{h}\right).
\end{aligned}$$

Let

$$P_1(z_i,z_l)=\frac{1}{2}\left[s_iK\left(\frac{x_i-x}{h}\right)+s_lK\left(\frac{x_l-x}{h}\right)\right]\frac{1}{h^{2k+1}}K\left(\frac{w_l-w_i}{h}\right).$$

Then equation (2.10.13) becomes

$$\frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{l=i+1}^{n}P_1(z_i,z_l). \tag{2.10.14}$$

Following Powell et al. (1989), we first verify that $E\left[P_1(z_i,z_l)^2\right]=o_p(n)$.

$$\begin{aligned}
& E\left[P_1(z_i,z_l)^2\right] \\
= & \int\int_{\Omega_{w_i,w_l}} E\left\{\left[\frac{1}{2}\left[s_iK\left(\frac{x_i-x}{h}\right)+s_lK\left(\frac{x_l-x}{h}\right)\right]\frac{1}{h^{2k+1}}K\left(\frac{w_l-w_i}{h}\right)\right]^2\middle| w_i,w_l\right\} \\
& f_w(w_i)f_w(w_l)\mathrm{d}w_i\mathrm{d}w_l. \\
= & \int\int_{\Omega_{u_i,u_l}} \frac{1}{h^{2k+1}}E\left\{\left[\frac{1}{2}\left[s_iK\left(u_i\right)+s_lK\left(u_i+hu_l\right)\right]K(u_l)\right]^2\middle|(x+hu_i,v_i),(x+hu_i+hu_l,v_i+hu_l)\right\} \\
& f_w(x+hu_i,v_i)f_w\left(x+hu_i+hu_l,v_i+hu_l\right)\mathrm{d}u_i\mathrm{d}v_i\mathrm{d}u_l \\
= & O_p\left(\frac{1}{h^{2k+1}}\right)=o_p(n),
\end{aligned}$$

where the second equality holds by the change of variables $u_l=\frac{w_l-w_i}{h}$, $u_i=\frac{x_i-x}{h}$, the third equality holds by the bounds conditions, and the last equality holds by the assumption that $nh^{2k+1}\to\infty$. According to Lemma 3.2 in Powell et al. (1989), equation (2.10.14) is equal to

$$E\left[P_1(z_i,z_l)\right]+\frac{2}{n}\sum_{i=1}^{n}\left\{E\left[P_1(z_i,z_l)|z_i\right]-E\left[P_1(z_i,z_l)\right]\right\}+o_p\left(\frac{1}{\sqrt{n}}\right). \tag{2.10.15}$$

The term inside the summation in equation (2.10.15) has the following form:

$$\begin{aligned}
& E\left[P_1(z_i,z_l)|z_i\right]-E\left[P_1(z_i,z_l)\right] \\
= & \frac{1}{2}E\left[\frac{s_i}{h^{2k+1}}K\left(\frac{x_i-x}{h}\right)K\left(\frac{w_i-w_l}{h}\right)\middle| z_i\right]+\frac{1}{2}E\left[\frac{s_l}{h^{2k+1}}K\left(\frac{x_l-x}{h}\right)K\left(\frac{w_i-w_l}{h}\right)\middle| z_i\right] \\
& -E\left[\frac{s_i}{h^{2k+1}}K\left(\frac{x_i-x}{h}\right)K\left(\frac{w_i-w_l}{h}\right)\right].
\end{aligned}$$

Since

$$\begin{aligned}
& E\left[\frac{s_i}{h^{2k+1}}K\left(\frac{x_i-x}{h}\right)K\left(\frac{w_i-w_l}{h}\right)\middle| z_i\right] \\
= & \int_{\Omega_{w_l}}\frac{s_i}{h^{2k+1}}K\left(\frac{x_i-x}{h}\right)K\left(\frac{w_i-w_l}{h}\right)f_w(w_l)\mathrm{d}w_l \\
= & \frac{s_i}{h^k}K\left(\frac{x_i-x}{h}\right)f_w(w_i)+\frac{s_i}{h^k}K\left(\frac{x_i-x}{h}\right)\int_{\Omega_{u_l}}K\left(u_l\right)\left[f_w(w_i+hu_l)-f_w(w_i)\right]\mathrm{d}u_l,
\end{aligned}$$

and similarly

$$\begin{aligned}
& E\left[\frac{s_l}{h^{2k+1}}K\left(\frac{x_l-x}{h}\right)K\left(\frac{w_i-w_l}{h}\right)\middle| z_i\right] \\
= & \frac{E\left[s_i|w_i\right]}{h^k}K\left(\frac{x_i-x}{h}\right)f_w(w_i)+\frac{1}{h^k}\int_{\Omega_{u_l}}\left[E\left[s_i|w_i+hu_l\right]K\left(\frac{x_i+hu_l-x}{h}\right)f_w(w_i+hu_l)\right. \\
& \left.-E\left[s_i|w_i\right]K\left(\frac{x_i-x}{h}\right)f_w(w_i)\right]K\left(u_l\right)\mathrm{d}u_l,
\end{aligned}$$

the following holds

$$\begin{aligned}
& E\left[P_1(z_i,z_l)|z_i\right]-E\left[P_1(z_i,z_l)\right] \\
= & \frac{1}{2}\frac{s_i}{h^k}K\left(\frac{x_i-x}{h}\right)f_w(w_i)+\frac{1}{2}\frac{E\left[s_i|w_i\right]}{h^k}K\left(\frac{x_i-x}{h}\right)f_w(w_i)-E\left[\frac{s_i}{h^k}K\left(\frac{x_i-x}{h}\right)f_w(w_i)\right] \\
& +R_{1i}-E\left(R_{1i}\right), \qquad (2.10.16)
\end{aligned}$$

where

$$\begin{aligned}
R_{1i} = & \frac{s_i}{h^k}K\left(\frac{x_i-x}{h}\right)\int K\left(u_l\right)\left[f_w(w_i+hu_l)-f_w(w_i)\right]\mathrm{d}u_l \qquad (2.10.17) \\
& +\frac{1}{h^k}\int_{\Omega_{u_l}}\left[E\left[s_i|w_i+hu_l\right]K\left(\frac{x_i+hu_l-x}{h}\right)f_w(w_i+hu_l)\right. \\
& \left.-E\left[s_i|w_i\right]K\left(\frac{x_i-x}{h}\right)f_w(w_i)K\left(u_l\right)\mathrm{d}u_l\right]
\end{aligned}$$

and, since $E(s_i|w_i)$, f_x, f_w satisfy the Lipschitz condition, $E\left(R_{1i}^2\right) = o_p\left(\frac{1}{h^k}\right)$. So

$$\frac{1}{n}\sum_{i=1}^{n}\left[R_{1i} - E\left(R_{1i}\right)\right] = o_p\left(\frac{1}{\sqrt{nh^k}}\right). \tag{2.10.18}$$

By the fact that $p(z_i, z_l)$ are symmetric for z_i, z_l, we have

$$E\left[P_1(z_i, z_l)\right] = E\left[s_i K\left(\frac{x_i - x}{h}\right) K(\frac{w_l - w_i}{h})\right].$$

From equation (2.10.15) (2.10.16) and (2.10.18), we have

$$\begin{aligned}
&\frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{l=i+1}^{n} P_1(z_i, z_l) \\
= \;& \frac{1}{n}\sum_{i=1}^{n}\left[\frac{s_i}{h^k}K\left(\frac{x_i - x}{h}\right) f_w\left(w_i\right) + \frac{E\left(s_i|w_i\right)}{h^k}K\left(\frac{x_i - x}{h}\right) f_w\left(w_i\right)\right. \\
& \left. -2E\left(\frac{s_i}{h^k}K\left(\frac{x_i - x}{h}\right) f_w\left(w_i\right)\right)\right] + E\left[s_i K\left(\frac{x_i - x}{h}\right) K(\frac{w_l - w_i}{h})\right] + o_P\left(\frac{1}{\sqrt{nh^k}}\right),
\end{aligned}$$

which implies the first part of the Theorem.

The second part holds by the same line of analysis by replacing W with X. ■

Lemma 2.10.3 *Adopt the same notation and assumptions as in Lemma 2.10.2, and assume $D^p f_x$ and $D^p f_w$ also satisfy the Lipschitz condition. Then*

$$\begin{aligned}
& E\left[\frac{s_i}{h^{2k+1}}K\left(\frac{x_i - x}{h}\right) K\left(\frac{w_l - w_i}{h}\right)\right] \\
= \;& E\left[\frac{s_i f_w\left(w_i\right)}{h^k}K\left(\frac{x_i - x}{h}\right)\right] + \mathbb{S}_{1,p} f_x\left(x\right) + o\left(h^p\right) \qquad (2.10.19) \\
& E\left[\frac{s_i}{h^{2k}}K\left(\frac{x_i - x}{h}\right) K\left(\frac{x_l - x_i}{h}\right)\right] \\
= \;& E\left[\frac{s_i f_x\left(x_i\right)}{h^k}K\left(\frac{x_i - x}{h}\right)\right] + \mathbb{S}_{2,p} f_x\left(x\right) + o\left(h^p\right) \qquad (2.10.20)
\end{aligned}$$

where

$$\begin{aligned}\mathbb{S}_{1,p} &\equiv h^p \sum_{|\mathbf{m}_{k+1}|=p} \frac{E\left[s_i D^{\mathbf{m}_{k+1}} f_w(w_i)|\, x\right]}{\mathbf{m}_{k+1}!} \int_{\mathbb{R}^{k+1}} u_l^{\mathbf{m}_{k+1}} K(u_l)\, du_l, \\ \mathbb{S}_{2,p} &\equiv h^p \sum_{|\mathbf{m}_k|=p} \frac{E\left[s_i D^{\mathbf{m}_k} f_w(x_i)|\, x\right]}{\mathbf{m}_k!} \int_{\mathbb{R}^k} u_l^{\mathbf{m}_k} K(u_l)\, du_l.\end{aligned}$$

Proof of Lemma 2.10.3.2

$$\begin{aligned}& E\left[\frac{s_i}{h^{2k+1}} K\left(\frac{x_i - x}{h}\right) K\left(\frac{w_l - w_i}{h}\right)\right] \\ = & \int_{\Omega_{\omega_i}} \int_{\Omega_{\omega_l}} \frac{1}{h^{k+1}} K\left(\frac{w_l - w_i}{h}\right) f_w(w_l)\, \mathrm{d}w_l E\left[\left.\frac{s_i}{h^k} K\left(\frac{x_i - x}{h}\right)\right| w_i\right] f_w(w_i)\, \mathrm{d}w_i \\ = & B_1 + E\left[\frac{s_i f_w(w_i)}{h^k} K\left(\frac{x_i - x}{h}\right)\right],\end{aligned}$$

where

$$B_1 \equiv \int_{\Omega_{\omega_i}} \int_{\Omega_{\omega_l}} \frac{1}{h^{k+1}} K\left(\frac{w_l - w_i}{h}\right) (f_w(w_l) - f_w(w_i))\, \mathrm{d}w_l E\left[\left.\frac{s_i}{h^k} K\left(\frac{x_i - x}{h}\right)\right| w_i\right] f_w(w_i)\, \mathrm{d}w_i.$$

Then, doing the standard change of variables transformation $u_l = \frac{w_l - w_i}{h}$, we have

$$B_1 = h^p \sum_{|\mathbf{m}_{k+1}|=p} \int_{\Omega_{\omega_i}} \int_{\mathbb{R}^{k+1}} u_l^{\mathbf{m}_{k+1}} K(u_l) \frac{D^{\mathbf{m}_{k+1}} f_w(\widetilde{w}_i)}{\mathbf{m}_{k+1}!} du_l E\left[\left.\frac{s_i}{h^k} K\left(\frac{x_i - x}{h}\right)\right| w_i\right] f_w(w_i)\, \mathrm{d}w_i,$$

where $\widetilde{w}_i$ is some value between w_i and $w_i + hu_l$. Since the kernel has bounded support and $D^p f_w(\widetilde{w}_i)$ satisfies the Lipschitz condition, we have

$$\begin{aligned}B_1 &= h^p \sum_{|\mathbf{m}_{k+1}|=p} \int_{\mathbb{R}^{k+1}} u_l^{\mathbf{m}_{k+1}} K(u_l)\, du_l \int_{\Omega_{\omega_i}} \frac{E\left[\left.\frac{s_i}{h^k} K\left(\frac{x_i - x}{h}\right) D^{\mathbf{m}_{k+1}} f_w(w_i)\right| w_i\right]}{\mathbf{m}_{k+1}!} f_w(w_i)\, \mathrm{d}w_i + o(h^p) \\ &= h^p \sum_{|\mathbf{m}_{k+1}|=p} \int_{\mathbb{R}^{k+1}} u_l^{\mathbf{m}_{k+1}} K(u_l)\, du_l \frac{E\left[\frac{s_i}{h^k} K\left(\frac{x_i - x}{h}\right) D^{\mathbf{m}_{k+1}} f_w(w_i)\right]}{\mathbf{m}_{k+1}!} + o(h^p).\end{aligned}$$

Substituting this into B_1, we have

$$B_1 = \mathbb{S}_{1,p} + o\left(h^p\right),$$

which is equation (2.10.19). The second conclusion can be proved similarly. ■

Corollary 2.10.4 *Under the same assumptions in Lemma 2.10.3 and assuming* $E\left(s_i|x\right)$ *and* $E\left(s_i|w\right)$ *are p-th order differentiable, with bounded p-th order derivatives, we have*

$$\widehat{E}\left[\left.\frac{s_i}{\widehat{f}(v_i|x_i)}\right| x\right]\widehat{f}_x(x) - E\left[\left.\frac{s_i}{f(v_i|x_i)}\right| x\right] f_x(x) = O_p(h^p) + O_p\left(\frac{1}{\sqrt{nh^k}}\right) + O_P\left(\frac{\log\left(n\right)}{nh^{k+1}}\right). \tag{2.10.21}$$

Proof of Corollary 2.10.4.2

$$\begin{aligned} & \widehat{E}\left[\left.\frac{s_i}{\widehat{f}(v_i|x_i)}\right| x\right]\widehat{f}_x(x) - E\left[\left.\frac{s_i}{f(v_i|x_i)}\right| x\right] f_x(x) \\ = \; & \widehat{E}\left[\left.\frac{s_i}{\widehat{f}(v_i|x_i)}\right| x\right]\widehat{f}_x(x) - \widehat{E}\left[\left.\frac{s_i}{f(v_i|x_i)}\right| x\right]\widehat{f}_x(x) + \widehat{E}\left[\left.\frac{s_i}{f(v_i|x_i)}\right| x\right]\left[\widehat{f}_x(x) - f_x(x)\right] \\ & + \left\{\widehat{E}\left[\left.\frac{s_i}{f(v_i|x_i)}\right| x\right] - E\left[\left.\frac{s_i}{f(v_i|x_i)}\right| x\right]\right\} f_x(x). \end{aligned}$$

All terms except the first term are readily seen to be $O_p(h^p) + O_p\left(\frac{1}{\sqrt{nh^k}}\right)$. For the first

term

$$
\begin{aligned}
& \widehat{E}\left[\left.\frac{s_i}{\widehat{f}(v_i|x_i)}\right| x\right]\widehat{f}_x(x) - \widehat{E}\left[\left.\frac{s_i}{f(v_i|x_i)}\right| x\right]\widehat{f}_x(x) \\
= & \frac{1}{n}\sum_{i=1}^{n}\frac{s_i\widehat{f}_x(x_i)}{\widehat{f}_w(w_i)}\frac{1}{h^k}K\left(\frac{x_i-x}{h}\right) - \frac{1}{n}\sum_{i=1}^{n}\frac{s_i f_x(x_i)}{f_w(w_i)}\frac{1}{h^k}K\left(\frac{x_i-x}{h}\right) \\
= & \frac{1}{n}\sum_{i=1}^{n}\frac{s_i\left[\widehat{f}_x(x_i)-f_x(x_i)\right]}{f_w(w_i)}\frac{1}{h^k}K\left(\frac{x_i-x}{h}\right) \qquad (2.10.22) \\
& +\frac{1}{n}\sum_{i=1}^{n}\frac{s_i f_x(x_i)\left[\widehat{f}_w(w_i)-f_w(w_i)\right]}{f_w^2(w_i)}\frac{1}{h^k}K\left(\frac{x_i-x}{h}\right) \qquad (2.10.23) \\
& +\frac{1}{n}\sum_{i=1}^{n}\frac{s_i\left[\widehat{f}_x(x_i)-f_x(x_i)\right]\left[\widehat{f}_w(w_i)-f_w(w_i)\right]}{f_w^2(w_i)}\frac{1}{h^k}K\left(\frac{x_i-x}{h}\right) \qquad (2.10.24) \\
& +\frac{1}{n}\sum_{i=1}^{n}\frac{s_i\widehat{f}_x(x_i)\left[\widehat{f}_w(w_i)-f_w(w_i)\right]^2}{f_w^2(w_i)\widehat{f}_w(w_i)}\frac{1}{h^k}K\left(\frac{x_i-x}{h}\right). \qquad (2.10.25)
\end{aligned}
$$

According the results in Lemma 2.10.2 and Lemma 2.10.3, equation (2.10.22) and (2.10.23) are $O_p(h^p) + O_p\left(\frac{1}{\sqrt{nh^k}}\right)$. From Silverman (1978) and Remark 2.10.1, we have

$$
\sup_{K\left(\frac{x_i-x}{h}\right)\neq 0,}\left|\widehat{f}_x(x_i)-f_x(x_i)\right| = O_p\left[\sqrt{\frac{\log(n)}{nh^k}}\right], \qquad (2.10.26)
$$

$$
\sup_{K\left(\frac{x_i-x}{h}\right)\neq 0,\ I_{\tau i}\neq 0}\left|\widehat{f}_w(w_i)-f_w(w_i)\right| = O_p\left[\sqrt{\frac{\log(n)}{nh^{k+1}}}\right]. \qquad (2.10.27)
$$

Then

$$
\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{n}\frac{s_i\left[\widehat{f}_x(x_i)-f_x(x_i)\right]\left[\widehat{f}_w(w_i)-f_w(w_i)\right]}{f_w^2(w_i)}\frac{1}{h^k}K\left(\frac{x_i-x}{h}\right)\right| \\
\leq & \sup_{K\left(\frac{x_i-x}{h}\right)\neq 0,\ I_{\tau i}\neq 0}\left|\widehat{f}_x(x_i)-f_x(x_i)\right|\left|\widehat{f}_w(w_i)-f_w(w_i)\right|\frac{1}{n}\sum_{i=1}^{n}\left|\frac{s_i}{f_w^2(w_i)h^k}K\left(\frac{x_i-x}{h}\right)\right| \\
= & O_P\left(\frac{\log(n)}{nh^{k+1/2}}\right),
\end{aligned}
$$

and similarly, we have

$$
\frac{1}{n}\sum_{i=1}^{n}\frac{s_i\widehat{f}_x(x_i)\left[\widehat{f}_w(w_i)-f_w(w_i)\right]^2}{f_w^2(w_i)\widehat{f}_w(w_i)}\frac{1}{h^k}K\left(\frac{x_i-x}{h}\right)=O_P\left(\frac{\log(n)}{nh^{k+1}}\right)
$$

Therefore, we know that equation (2.10.24) and (2.10.25) are of the orders $O_P\left(\frac{\log(n)}{nh^{k+1/2}}\right)$ and $O_P\left(\frac{\log(n)}{nh^{k+1}}\right)$ respectively.

Combining the above results the proves the Corollary. ■

Proof of Theorem 2.3.5.2 We first derive the properties of $\widehat{\psi}_1(x)$. This can be divided into several components as follows

$$
\begin{aligned}
\widehat{\psi}_1(x)-\psi_1(x) &= \frac{\widehat{E}\left(\widehat{h}_{1i}\middle|x\right)}{\widehat{E}\left(\widehat{g}_{1i}|x\right)}=\frac{\widehat{E}\left(\widehat{\widetilde{h}}_{1i}\middle|x\right)}{\widehat{E}\left(\widehat{\widetilde{g}}_{1i}\middle|x\right)} \\
&= \frac{\widehat{E}\left(\widehat{\widetilde{h}}_{1i}\middle|x\right)}{E\left(\widetilde{g}_{1i}|x\right)}-\frac{E\left(\widetilde{h}_{1i}\middle|x\right)\widehat{E}\left(\widehat{\widetilde{g}}_{1i}\middle|x\right)}{E\left(\widetilde{g}_{1i}|x\right)^2}+R_2(x), \qquad (2.10.28)
\end{aligned}
$$

where

$$
R_2(x)\equiv\frac{\left[\widehat{E}\left(\widehat{\widetilde{h}}_{1i}\middle|x\right)-E\left(\widetilde{h}_{1i}\middle|x\right)\right]\left[\widehat{E}\left(\widehat{\widetilde{g}}_{1i}\middle|x\right)-E\left(\widetilde{g}_{1i}|x\right)\right]}{\left[E\left(\widetilde{g}_{1i}|x\right)\right]^2}+\frac{\widehat{E}\left(\widehat{\widetilde{h}}_{1i}\middle|x\right)\left[\widehat{E}\left(\widehat{\widetilde{g}}_{1i}\middle|x\right)-E\left(\widetilde{g}_{1i}|x\right)\right]^2}{\left[E\left(\widetilde{g}_{1i}|x\right)\right]^2}.
$$

According to Corollary 2.10.4, and the assumption that $\frac{1}{E(\widetilde{g}_{1i}|x)}$ is bounded, $R_2(x)$ is of order $o_P\left(\frac{1}{\sqrt{nh^k}}\right)$. So

$$\widehat{\psi}_1(x) - \psi_1(x) = \frac{\widehat{E}\left(\widehat{\widetilde{h}}_{1i}\middle| x\right)}{E\left(\widetilde{g}_{1i}|\, x\right)} - \frac{E\left(\widetilde{h}_{1i}\middle| x\right)\widehat{E}\left(\widehat{\widetilde{g}}_{1i}\middle| x\right)}{\left[E\left(\widetilde{g}_{1i}|\, x\right)\right]^2} + o_P\left(\frac{1}{\sqrt{nh^k}}\right). \tag{2.10.29}$$

Notice that

$$\begin{aligned}
\frac{\widehat{E}\left(\widehat{\widetilde{h}}_{1i}\middle| x\right)}{E\left(\widetilde{g}_{1i}|\, x\right)} &= \frac{1}{E\left(\widetilde{g}_{1i}|\, x\right)}\frac{1}{n}\sum_{i=1}^{n}\frac{D_iY_i}{\widehat{f}(v_i|x_i)}\frac{1}{h^k}K\left(\frac{x_i-x}{h}\right) \qquad (2.10.30)\\
= \quad & \frac{1}{E\left(\widetilde{g}_{1i}|\, x\right)}\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{D_iY_i\widehat{f}_x(x_i)}{f_{xv}(x_i,v_i)}\frac{1}{h^k}K\left(\frac{x_i-x}{h}\right)\right.\\
& \left.-\frac{D_iY_if_x(x_i)\left[\widehat{f}_{xv}(x_i,v_i)-f_{xv}(x_i,v_i)\right]}{f_{xv}^2(x_i,v_i)}\frac{1}{h^k}K\left(\frac{x_i-x}{h}\right)+R_{3i}\right\},
\end{aligned}$$

where

$$\begin{aligned}
R_{3i} \quad \equiv \quad & \frac{D_iY_i\widehat{f}_x(x_i)\left[\widehat{f}_{xv}(x_i,v_i)-f_{xv}(x_i,v_i)\right]^2}{E\left(\widetilde{g}_{1i}|\, x\right)f_{xv}^2(x_i,v_i)\widehat{f}_{xv}(x_i,v_i)}\frac{1}{h^k}K\left(\frac{x_i-x}{h}\right)\\
& -\frac{D_iY_i\left[\widehat{f}_x(x_i)-f_x(x_i)\right]\left[\widehat{f}_{xv}(x_i,v_i)-f_{xv}(x_i,v_i)\right]}{E\left(\widetilde{g}_{1i}|\, x\right)f_{xv}^2(x_i,v_i)}\frac{1}{h^k}K\left(\frac{x_i-x}{h}\right).
\end{aligned}$$

Following the same proof in Corollary 2.10.4

$$\frac{1}{n}\sum_{i=1}^{n}R_{3i} = O_p\left(\frac{\log(n)}{nh^{2k+1}}\right) = o_p\left(\frac{1}{\sqrt{nh^k}}\right). \tag{2.10.31}$$

Apply Lemma 2.10.2 on the first term in equation (2.10.30),

$$\begin{aligned} & \frac{1}{E\left(\widetilde{g}_{1i}|\,x\right)}\frac{1}{n}\sum_{i=1}^{n}\frac{D_iY_i\widehat{f}_x(x_i)}{f_{xv}(x_i,v_i)}\frac{1}{h^k}K\left(\frac{x_i-x}{h}\right) \qquad (2.10.32)\\ = & \frac{1}{E\left(\widetilde{g}_{1i}|\,x\right)}\frac{1}{n}\sum_{i=1}^{n}\left[\frac{h_{1i}}{h^k}K\left(\frac{x_i-x}{h}\right)+\frac{E\left(h_{1i}|x_i\right)}{h^k}K\left(\frac{x_i-x}{h}\right)\right.\\ & \left.-2E\left(\frac{h_{1i}}{h^k}K\left(\frac{x_i-x}{h}\right)\right)\right]+E\left[\frac{D_iY_i}{f_{xv}(x_i,v_i)}\frac{1}{h^{2k}}K\left(\frac{x_i-x}{h}\right)K\left(\frac{x_l-x_i}{h}\right)\right]. \end{aligned}$$

By the same reasoning, the second component in equation (2.10.30) is

$$\begin{aligned} & \frac{1}{E\left(\widetilde{g}_{1i}|\,x\right)}\frac{1}{n}\sum_{i=1}^{n}\frac{D_iY_if_x(x_i)\widehat{f}_{xv}(x_i,v_i)}{f_{xv}^2(x_i,v_i)}\frac{1}{h^k}K\left(\frac{x_i-x}{h}\right) \qquad (2.10.33)\\ = & \frac{1}{E\left(\widetilde{g}_{1i}|\,x\right)}\frac{1}{n}\sum_{i=1}^{n}\left[\frac{h_{1i}}{h^k}K\left(\frac{x_i-x}{h}\right)+\frac{E\left(h_{1i}|x_i,v_i\right)}{h^k}K\left(\frac{x_i-x}{h}\right)\right.\\ & \left.-2E\left(\frac{h_{1i}}{h^k}K\left(\frac{x_i-x}{h}\right)\right)\right]+E\left[\frac{D_iY_if_x(x_i)}{f_{xv}^2(x_i,v_i)}\frac{1}{h^{2k}}K\left(\frac{x_i-x}{h}\right)K\left(\frac{w_l-w_i}{h}\right)\right]. \end{aligned}$$

Substituting equation (2.10.32) and (2.10.33) back into equation (2.10.30) and using the results in Lemma 2.10.3, we have

$$\begin{aligned} \frac{\widehat{E}\left(\widehat{\widetilde{h}}_{1i}\middle|\,x\right)}{E\left(\widetilde{g}_{1i}|\,x\right)} = & \frac{1}{E\left(\widetilde{g}_{1i}|\,x\right)}\frac{1}{n}\sum_{i=1}^{n}\left[h_{1i}+E\left(h_{1i}|x_i\right)-E\left(h_{1i}|x_i,v_i\right)\right]\frac{1}{h^k}K\left(\frac{x_i-x}{h}\right) \qquad (2.10.34)\\ & +\frac{\mathbb{B}_{1,p}}{E\left(g_{1i}|\,x\right)}-\frac{\mathbb{B}_{2,p}}{E\left(g_{1i}|\,x\right)}+o_P\left(h^p\right). \end{aligned}$$

Applying the same strategy to the next term in equation (2.10.29), we get

$$\begin{aligned} \frac{E\left(\widetilde{h}_{1i}\middle|\,x\right)\widehat{E}\left(\widehat{\widetilde{g}}_{1i}\middle|\,x\right)}{\left[E\left(\widetilde{g}_{1i}|\,x\right)\right]^2} = & \frac{E\left(\widetilde{h}_{1i}\middle|\,x\right)}{E\left(\widetilde{g}_{1i}|\,x\right)^2}\frac{1}{n}\sum_{i=1}^{n}\left[g_{1i}+E\left(g_{1i}|x_i\right)-E\left(g_{1i}|x_i,v_i\right)\right]\frac{1}{h^k}K\left(\frac{x_i-x}{h}\right) \qquad (2.10.35)\\ & +\frac{E\left(h_{1i}|\,x\right)\mathbb{B}_{3,p}}{E\left(g_{1i}|\,x\right)^2}-\frac{E\left(h_{1i}|\,x\right)\mathbb{B}_{4,p}}{E\left(g_{1i}|\,x\right)^2}+o_P\left(h^p\right) \end{aligned}$$

Substituting equation (2.10.34) and (2.10.35) into equation (2.10.29), we have

$$\begin{aligned}\widehat{\psi}_1(x)-\psi_1(x) &= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{h_{1i}}{E\left(\widetilde{g}_{1i}\middle|x\right)}+\frac{E\left(h_{1i}|x_i\right)}{E\left(\widetilde{g}_{1i}\middle|x\right)}-\frac{E\left(h_{1i}|x_i,v_i\right)}{E\left(\widetilde{g}_{1i}\middle|x\right)}-\frac{E\left(\widetilde{h}_{1i}\middle|x\right)g_{1i}}{E\left(\widetilde{g}_{1i}\middle|x\right)^2}\right.\\ &\quad\left.-\frac{E\left(\widetilde{h}_{1i}\middle|x\right)E\left(g_{1i}|x_i\right)}{E\left(\widetilde{g}_{1i}\middle|x\right)^2}+\frac{E\left(\widetilde{h}_{1i}\middle|x\right)E\left(g_{1i}|x_i,v_i\right)}{E\left(\widetilde{g}_{1i}\middle|x\right)^2}\right]\frac{1}{h^k}K\left(\frac{x_i-x}{h}\right)\\ &\quad+\frac{\mathbb{B}_{1,p}}{E\left(g_{1i}\middle|x\right)}-\frac{\mathbb{B}_{2,p}}{E\left(g_{1i}\middle|x\right)}-\frac{E\left(h_{1i}\middle|x\right)\mathbb{B}_{3,p}}{E\left(g_{1i}\middle|x\right)^2}+\frac{E\left(h_{1i}\middle|x\right)\mathbb{B}_{4,p}}{E\left(g_{1i}\middle|x\right)^2}+o_P\left(h^p\right)o_p\left(\frac{1}{\sqrt{nh^k}}\right).\end{aligned}$$

Similarly,

$$\begin{aligned}\widehat{\psi}_2(x)-\psi_2(x) &= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{h_{2i}}{E\left(\widetilde{g}_{2i}\middle|x\right)}+\frac{E\left(h_{2i}|x_i\right)}{E\left(\widetilde{g}_{2i}\middle|x\right)}-\frac{E\left(h_{2i}|x_i,v_i\right)}{E\left(\widetilde{g}_{2i}\middle|x\right)}-\frac{E\left(\widetilde{h}_{2i}\middle|x\right)g_{2i}}{E\left(\widetilde{g}_{2i}\middle|x\right)^2}\right.\\ &\quad\left.-\frac{E\left(\widetilde{h}_{2i}\middle|x\right)E\left(g_{2i}|x_i\right)}{E\left(\widetilde{g}_{2i}\middle|x\right)^2}+\frac{E\left(\widetilde{h}_{2i}\middle|x\right)E\left(g_{2i}|x_i,v_i\right)}{E\left(\widetilde{g}_{2i}\middle|x\right)^2}\right]\frac{1}{h^k}K\left(\frac{x_i-x}{h}\right)\\ &\quad+\frac{\mathbb{B}_{5,p}}{E\left(g_{2i}\middle|x\right)}-\frac{\mathbb{B}_{6,p}}{E\left(g_{2i}\middle|x\right)}-\frac{E\left(h_{2i}\middle|x\right)\mathbb{B}_{7,p}}{E\left(g_{2i}\middle|x\right)^2}+\frac{E\left(h_{2i}\middle|x\right)\mathbb{B}_{8,p}}{E\left(g_{2i}\middle|x\right)^2}+o_P\left(h^p\right)+o_p\left(\frac{1}{\sqrt{nh^k}}\right).\end{aligned}$$

Putting these results together gives

$$\widehat{\psi}_1(x)-\widehat{\psi}_2(x)-(\psi_1(x)-\psi_2(x))=\frac{1}{n}\sum_{i=1}^{n}q_i\left(x\right)\frac{1}{h^k}K\left(\frac{x_i-x}{h}\right)+\mathbb{B}_p\left(x\right)+o_P\left(h^p\right)+o_p\left(\frac{1}{\sqrt{nh^k}}\right),$$

which implies that

$$\frac{\sqrt{nh^k}}{var\left(q_i\left(x\right)|x\right)\int_{\mathbb{R}^k}K^2\left(u\right)\mathrm{d}u}\left[\widehat{\psi}_1(x)-\widehat{\psi}_2(x)-(\psi_1(x)-\psi_2(x))-\mathbb{B}_p\left(x\right)\right]\xrightarrow{d}N\left(0,1\right)$$

■

Proof of Theorem 2.3.9.2 The first-order asymptotics of our estimator follow directly from of Lemmas 2.10.5, 2.10.8, 2.10.9 and 2.10.10. The convergence rate of the resulting

influence function can be seen from Lemmas 2.10.8, 2.10.9 and 2.10.10. ■

Lemma 2.10.5 *Let Assumptions 21, 22, 23, 24, 25, 26, 37, and 39 hold. Assume that bandwidth* $h = c_0 n^{-c_T/2}$ *in* $\widehat{f}_{v_t}$*, and assume a kernel of order* $p \geq (1 - c_T/2)/c_T$*. Then*

$$\begin{aligned}
&\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} D_{it}Y_{it}\Big/\widehat{f}_{v_t}(v_{it})}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} D_{it}\Big/\widehat{f}_{v_t}(v_{it})} - \frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}(1-D_{it})Y_{it}\Big/\widehat{f}_{v_t}(v_{it})}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}(1-D_{it})\Big/\widehat{f}_{v_t}(v_{it})} - [E(Y_1) - E(Y_0)] \\
= &\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\Lambda_{1it}}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\Pi_{1it}} - \frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\Lambda_{2it}}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\Pi_{2it}} + o_P\left(\frac{1}{\sqrt{nT}}\right).
\end{aligned}$$

Proof of Lemma 2.10.5.2 First note that

$$\begin{aligned}
&\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} D_{it}Y_{it}\Big/\widehat{f}_{v_t}(v_{it})}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} D_{it}\Big/\widehat{f}_{v_t}(v_{it})} - \frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}(1-D_{it})Y_{it}\Big/\widehat{f}_{v_t}(v_{it})}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}(1-D_{it})\Big/\widehat{f}_{v_t}(v_{it})} - [E(Y_1) - E(Y_0)] \\
= &\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} D_{it}\left(Y_{it} - E\left(\widetilde{a}_i + \widetilde{b}_t + Y_1\right)\right)\Big/\widehat{f}_{v_t}(v_{it})}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} D_{it}\Big/\widehat{f}_{v_t}(v_{it})} \\
&- \frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}(1-D_{it})\left(Y_{it} - E\left(\widetilde{a}_i + \widetilde{b}_t + Y_1\right)\right)\Big/\widehat{f}_{v_t}(v_{it})}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}(1-D_{it})\Big/\widehat{f}_{v_t}(v_{it})}.
\end{aligned}$$

We first show that

$$\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{D_{it}\left(Y_{it} - E(\widetilde{a}_i + \widetilde{b}_t + Y_1)\right)}{\widehat{f}_{v_t}(v_{it})} = \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\Lambda_{1it} + o_p(\frac{1}{\sqrt{nT}}).$$

To this end,

$$\begin{aligned} & \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{D_{it}\left(Y_{it}-E(\widetilde{a}_i+\widetilde{b}_t+Y_1)\right)}{\widehat{f}_{v_t}(v_{it})} \\ = \; & \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{D_{it}\left(Y_{it}-E(\widetilde{a}_i+\widetilde{b}_t+Y_1)\right)}{f_{v_t}(v_{it})} \\ & -\frac{D_{it}\left(Y_{it}-E(\widetilde{a}_i+\widetilde{b}_t+Y_1)\right)}{f_{v_t}^2(v_{it})}\left(\widehat{f}_{v_t}(v_{it})-f_{v_t}(v_{it})\right)+R_{nit}, \end{aligned} \tag{2.10.36}$$

where

$$R_{nit}\equiv\frac{D_{it}\left(Y_{it}-E(\widetilde{a}_i+\widetilde{b}_t+Y_1)\right)}{f_{v_t}^2(v_{it})\widehat{f}_{v_t}(v_{it})}\left(\widehat{f}_{v_t}(v_{it})-f_{v_t}(v_{it})\right)^2.$$

Again, by the uniform convergence of $\widehat{f}_{v_t}(v_{it})$ (our assumption on p guarantees that the bias term vanishes fast enough),

$$\sup_{I_{\tau it}\neq 0}\left|\widehat{f}_{v_t}\left(v_{it}\right)-f_v\left(v_{it}\right)\right|=O_P\left(\log\left(n\right)\Big/\sqrt{nh}\right)=O_P\left(\log\left(n\right)\Big/n^{1/2-c_T/4}\right)=o_p\left((nT)^{-1/4}\right),$$

such that $\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}|R_{nit}|=o_p\left(\frac{1}{\sqrt{nT}}\right)$.

Generalizing Lemma 2.10.2 a little, we have, $E\left[p\left(z_i,z_j\right)^2\right]=O\left(1/h\right)=o\left(n/T\right)$, and

$$\begin{aligned} & \frac{1}{n}\sum_{i=1}^{n}\frac{D_{it}\left(Y_{it}-E(\widetilde{a}_i+\widetilde{b}_t+Y_1)\right)}{f_{v_t}^2(v_{it})}\left(\widehat{f}_{v_t}(v_{it})-f_{v_t}(v_{it})\right) \\ = \; & \frac{1}{n}\sum_{i=1}^{n}\frac{E\left[\left(Y_{it}-E(\widetilde{a}_i+\widetilde{b}_t+Y_1)\right)D_{it}\Big|\,v_{it}\right]}{f_{v_t}(v_{it})}+o_p\left(\frac{1}{\sqrt{nT}}\right), \end{aligned}$$

for $t=1,...,T$. Substitute this back into equation (2.10.36), we get that $\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{D_{it}\left(Y_{it}-E(\widetilde{a}_i+\widetilde{b}_t+Y_1)\right)}{\widehat{f}_{v_t}(v_{it})}$

is equal to

$$\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{\left(Y_{it}-E(\widetilde{a}_i+\widetilde{b}_t+Y_1)\right)D_{it}-E\left[\left(Y_{it}-E(\widetilde{a}_i+\widetilde{b}_t+Y_1)\right)D_{it}|v_{it}\right]}{f_{v_t}(v_{it})}+o_p\left(\frac{1}{\sqrt{nT}}\right),$$

which is $\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\Lambda_{1it}+o_p\left(\frac{1}{\sqrt{nT}}\right)$.

For the same reason

$$\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{D_{it}}{\widehat{f}_{v_t}(v_{it})}=\overline{\Pi}_1+\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{D_{it}-E\left(D_{it}|v_{it}\right)}{f_{v_t}(v_{it})}+o_p\left(\frac{1}{\sqrt{nT}}\right).$$

By the independence assumption on V_{it} across i and t, we know $\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\left(\frac{D_{it}}{\widehat{f}_{v_t}(v_{it})}-\frac{D_{it}}{f_{v_t}(v_{it})}\right)=O_p\left(\frac{1}{\sqrt{nT}}\right)$. Therefore

$$\begin{aligned}
&\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}D_{it}\left(Y_{it}-E\left(\widetilde{a}_i+\widetilde{b}_t+Y_1\right)\right)\Big/\widehat{f}_{v_t}(v_{it})}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}D_{it}\Big/\widehat{f}_{v_t}(v_{it})}=\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\Lambda_{1it}}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}D_{it}\Big/\widehat{f}_{v_t}(v_{it})}+o_p\left(\frac{1}{\sqrt{nT}}\right)\\
=&\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\Lambda_{1it}}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\Pi_{1it}}-\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\Lambda_{1it}\left(\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\left(D_{it}\Big/\widehat{f}_{v_t}(v_{it})-D_{it}/f_{v_t}(v_{it})\right)\right)}{\left(\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\Pi_{1it}\right)\left(\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}D_{it}\Big/\widehat{f}_{v_t}(v_{it})\right)}+o_p\left(\frac{1}{\sqrt{nT}}\right)\\
=&\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\Lambda_{1it}}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\Pi_{1it}}+o_p\left(\frac{1}{\sqrt{nT}}\right),
\end{aligned}$$

where the last equality holds by $\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\Lambda_{1it}=o_P(1)$ and $\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\left(\frac{D_{it}}{\widehat{f}_{v_t}(v_{it})}-\frac{D_{it}}{f_{v_t}(v_{it})}\right)=O_p\left(\frac{1}{\sqrt{nT}}\right)$. Applying the same analysis to the second component of the estimator finishes the proof. ■

Lemma 2.10.6 *Let Assumption 21, 22, 23, 25 hold, then*

$$\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{D_{it}}{f_{v_t}(v_{it})}-\overline{\Pi}_1=O_P\left((nT)^{-1/2}\right),$$
$$\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{1-D_{it}}{f_{v_t}(v_{it})}-\overline{\Pi}_2=O_P\left((nT)^{-1/2}\right).$$

Proof of Lemma 2.10.6.2 Here we prove the first equality of the lemma, and the second follows by the same logic. Note that

$$\begin{aligned}
E\left(\frac{D_{it}}{f_{v_t}(v_{it})}\middle|a_i,\widetilde{a}_i\right) &= E\left(E\left(\frac{D_{it}}{f_{v_t}(v_{it})}\middle|a_i,b_t,u_{it}\right)\middle|a_i,\widetilde{a}_i\right)\\
&= E\left(\int\frac{I\left(0\leq a_i+b_t+v_{it}+u_{it}\leq\alpha\right)}{f_{v_t}(v_{it})}f_{v_t}\left(v_{it}|a_i,\widetilde{a}_i,b_t,u_{it}\right)dv_{it}\middle|a_i,\widetilde{a}_i\right)\\
&= E\left(\int I\left(0\leq a_i+b_t+v_{it}+u_{it}\leq\alpha\right)dv_{it}\middle|a_i,\widetilde{a}_i\right)\\
&= \alpha=\overline{\Pi}_1.
\end{aligned}$$

Similarly, we have

$$E\left(\frac{D_{it}}{f_{v_t}(v_{it})}\middle|b_t,\widetilde{b}_t\right)=\alpha=\overline{\Pi}_1.$$

By this result, we have

$$\begin{aligned}
&\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{D_{it}}{f_{v_t}(v_{it})}-\overline{\Pi}_1\\
=\ &\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\left(\frac{D_{it}}{f_{v_t}(v_{it})}-E\left(\frac{D_{it}}{f_{v_t}(v_{it})}\middle|a_i,a_i\right)-E\left(\frac{D_{it}}{f_{v_t}(v_{it})}\middle|b_t,\widetilde{b}_t\right)+\overline{\Pi}_1\right).
\end{aligned}$$

By the conditional independence assumption, we know the covariance of the above terms

for either different i or t is zero. So we have

$$\sqrt{nT}\left(\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{D_{it}}{f_{v_t}(v_{it})}-\overline{\Pi}_1\right)\xrightarrow{d}N\left(0,var\left(\Pi_{1it}\right)\right).$$

The second part of the theorem follows similarly. ■

Lemma 2.10.7 *Let Assumption 21, 22, 23, 25, for $j=0,1$*

$$\begin{aligned}
E\left[\left(\frac{D_{it}}{f_{v_t}(v_{it})\overline{\Pi}_1}-1\right)\varepsilon_{jit}\middle|\,a_i,\widetilde{a}_i\right] &= E\left[\left(\frac{1-D_{it}}{f_{v_t}(v_{it})\overline{\Pi}_2}-1\right)\varepsilon_{jit}\middle|\,a_i,\widetilde{a}_i\right]=0,\\
E\left[\left(\frac{D_{it}}{f_{v_t}(v_{it})\overline{\Pi}_1}-1\right)\varepsilon_{jit}\middle|\,b_t,\widetilde{b}_t\right] &= E\left[\left(\frac{1-D_{it}}{f_{v_t}(v_{it})\overline{\Pi}_2}-1\right)\varepsilon_{jit}\middle|\,b_t,\widetilde{b}_t\right]=0,\\
E\left[\frac{D_{it}}{f_{v_t}(v_{it})\overline{\Pi}_1}-1\middle|\,a_i,\widetilde{a}_i\right] &= E\left[\frac{1-D_{it}}{f_{v_t}(v_{it})\overline{\Pi}_2}-1\middle|\,a_i,\widetilde{a}_i\right]=0,\\
E\left[\frac{D_{it}}{f_{v_t}(v_{it})\overline{\Pi}_1}-1\middle|\,b_t,\widetilde{b}_t\right] &= E\left[\frac{1-D_{it}}{f_{v_t}(v_{it})\overline{\Pi}_2}-1\middle|\,b_t,\widetilde{b}_t\right]=0.
\end{aligned}$$

Proof of Lemma 2.10.7.2 Note that by the proof of Lemma 2.10.6

$$\begin{aligned}
E\left[\left(\frac{D_{it}}{f_{v_t}(v_{it})\overline{\Pi}_1}-1\right)\varepsilon_{jit}\middle|\,a_i,\widetilde{a}_i\right] &= E\left(\varepsilon_{jit}|\,a_i,\widetilde{a}_i\right)-E\left(\varepsilon_{jit}|\,a_i,\widetilde{a}_i\right)=0,\\
E\left[\left(\frac{D_{it}}{f_{v_t}(v_{it})\overline{\Pi}_1}-1\right)\varepsilon_{jit}\middle|\,b_t,\widetilde{b}_t\right] &= E\left(\varepsilon_{jit}|\,b_t,\widetilde{b}_t\right)-E\left(\varepsilon_{jit}|\,b_t,\widetilde{b}_t\right)=0,
\end{aligned}$$

for $j=0,1$. Others follow similarly. ■

Lemma 2.10.8 *Let Assumption 21, 22, 23, 25 hold, then*

$$\begin{aligned}
&\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}D_{it}\left(\widetilde{a}_i+\widetilde{b}_t-E\left(\widetilde{a}_i+\widetilde{b}_t\right)\right)\Big/f_{v_t}(v_{it})}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}D_{it}/f_{v_t}(v_{it})}-\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}(1-D_{it})\left(\widetilde{a}_i+\widetilde{b}_t-E\left(\widetilde{a}_i+\widetilde{b}_t\right)\right)\Big/f_{v_t}(v_{it})}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}(1-D_{it})/f_{v_t}(v_{it})}\\
=\;&\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\left[\left(\frac{D_{it}}{f_{v_t}(v_{it})\overline{\Pi}_1}-\frac{1-D_{it}}{f_{v_t}(v_{it})\overline{\Pi}_2}\right)\left(\widetilde{a}_i-E\left(\widetilde{a}_i\right)+\widetilde{b}_t-E\left(\widetilde{b}_t\right)\right)\right]+o_P\left((nT)^{-1/2}\right),
\end{aligned}$$

and $\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\left[\left(\frac{D_{it}}{f_{v_t}(v_{it})\overline{\Pi}_1}-\frac{1-D_{it}}{f_{v_t}(v_{it})\overline{\Pi}_2}\right)\left(\widetilde{a}_i-E\left(\widetilde{a}_i\right)+\widetilde{b}_t-E\left(\widetilde{b}_t\right)\right)\right]=O_p\left((nT)^{-1/2}\right).$

Proof of Lemma 2.10.8.2

$$\begin{aligned}
&\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}D_{it}\left(\widetilde{a}_i+\widetilde{b}_t\right)\Big/ f_{v_t}(v_{it})}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}D_{it}/f_{v_t}(v_{it})}-\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}(1-D_{it})\left(\widetilde{a}_i+\widetilde{b}_t\right)\Big/ f_{v_t}(v_{it})}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}(1-D_{it})/f_{v_t}(v_{it})}\\
=\;&\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}D_{it}\left(\widetilde{a}_i+\widetilde{b}_t\right)\Big/ f_{v_t}(v_{it})}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}D_{it}/f_{v_t}(v_{it})}-\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\left(\widetilde{a}_i+\widetilde{b}_t\right)\\
&-\left(\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}(1-D_{it})\left(\widetilde{a}_i+\widetilde{b}_t\right)\Big/ f_{v_t}(v_{it})}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}(1-D_{it})/\,f_{v_t}(v_{it})}-\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\left(\widetilde{a}_i+\widetilde{b}_t\right)\right).
\end{aligned}$$

We analyze the first term.

$$
\begin{aligned}
&\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} D_{it}\left(\widetilde{a}_i+\widetilde{b}_t\right)\Big/ f_{v_t}(v_{it})}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} D_{it}/f_{v_t}(v_{it})} - \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\left(\widetilde{a}_i+\widetilde{b}_t\right) \\
= \;& \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\left(\frac{D_{it}}{f_{v_t}(v_{it})\overline{\Pi}_1}-1\right)\left(\widetilde{a}_i+\widetilde{b}_t\right) - \frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} D_{it}\left(\widetilde{a}_i+\widetilde{b}_t\right)\Big/ f_{v_t}(v_{it})\left(\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} D_{it}/f_{v_t}(v_{it})-\overline{\Pi}_1\right.}{\left(\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} D_{it}/f_{v_t}(v_{it})\right)\overline{\Pi}_1} \\
= \;& \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\left(\frac{D_{it}}{f_{v_t}(v_{it})\overline{\Pi}_1}-1\right)\left(\widetilde{a}_i-E\left(\widetilde{a}_i\right)+\widetilde{b}_t-E\left(\widetilde{b}_t\right)\right)+\left(E\left(\widetilde{a}_i\right)+E\left(\widetilde{b}_t\right)\right)\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\left(\frac{D_{it}}{f_{v_t}(v_{it})\overline{\Pi}_1}-\right. \\
& -\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} D_{it}\left(\widetilde{a}_i+\widetilde{b}_t\right)\Big/ f_{v_t}(v_{it})\left(\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} D_{it}/f_{v_t}(v_{it})-\overline{\Pi}_1\right)}{\left(\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} D_{it}/f_{v_t}(v_{it})\right)\overline{\Pi}_1} \\
= \;& \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\left(\frac{D_{it}}{f_{v_t}(v_{it})\overline{\Pi}_1}-1\right)\left(\widetilde{a}_i-E\left(\widetilde{a}_i\right)+\widetilde{b}_t-E\left(\widetilde{b}_t\right)\right)+o_P\left((nT)^{-1/2}\right).
\end{aligned}
$$

So we have

$$
\begin{aligned}
&\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{D_{it}\left(\widetilde{a}_i+\widetilde{b}_t\right)}{f_{v_t}(v_{it})}}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{D_{it}}{f_{v_t}(v_{it})}} - \frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{(1-D_{it})\left(\widetilde{a}_i+\widetilde{b}_t\right)}{f_{v_t}(v_{it})}}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{(1-D_{it})}{f_{v_t}(v_{it})}} \\
= \;& \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\left[\frac{D_{it}}{f_{v_t}(v_{it})\overline{\Pi}_1}-\frac{1-D_{it}}{f_{v_t}(v_{it})\overline{\Pi}_2}\right]\left(\widetilde{a}_i-E\left(\widetilde{a}_i\right)+\widetilde{b}_t-E\left(\widetilde{b}_t\right)\right)+o_P\left((nT)^{-1/2}\right).
\end{aligned}
$$

The rate of the influence function above can be similarly seen from Lemma 2.10.7. ■

Lemma 2.10.9 *Let Assumption 21, 22, 23, 25, and 26 hold, then*

$$\begin{aligned}
&\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} E\left[\left(Y_{it}-E\left(\widetilde{a}_i+\widetilde{b}_t+Y_1\right)\right)D_{it}\middle| v_{it}\right]\Big/ f_{v_t}(v_{it})}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} D_{it}/f_{v_t}(v_{it})} \\
= \quad & \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{E\left[\left(Y_{it}-E\left(\widetilde{a}_i+\widetilde{b}_t+Y_1\right)\right)D_{it}\middle| v_{it}\right]}{\overline{\Pi}_1 f_{v_t}(v_{it})}+o_P\left((nT)^{-1/2}\right)
\end{aligned}$$

$$\begin{aligned}
&\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} E\left[\left(Y_{it}-E\left(\widetilde{a}_i+\widetilde{b}_t+Y_0\right)\right)(1-D_{it})\middle| v_{it}\right]\Big/ f_{v_t}(v_{it})}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} (1-D_{it})/f_{v_t}(v_{it})} \\
= \quad & \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{E\left[\left(Y_{it}-E\left(\widetilde{a}_i+\widetilde{b}_t+Y_0\right)\right)(1-D_{it})\middle| v_{it}\right]}{\overline{\Pi}_2 f_{v_t}(v_{it})}+o_P\left((nT)^{-1/2}\right)
\end{aligned}$$

and $\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{E\left[\left(Y_{it}-E\left(\widetilde{a}_i+\widetilde{b}_t+Y_1\right)\right)D_{it}\middle| v_{it}\right]}{\overline{\Pi}_1 f_{v_t}(v_{it})}=O_p\left((nT)^{-1/2}\right)$, $\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{E\left[\left(Y_{it}-E\left(\widetilde{a}_i+\widetilde{b}_t+Y_0\right)\right)(1-D_{it})\middle| v_{it}\right]}{\overline{\Pi}_2 f_{v_t}(v_{it})}=O_p\left((nT)^{-1/2}\right)$.

Proof of Lemma 2.10.9.2 The first part of this theorem follows the same line proof as Lemma 2.10.8. The $\sqrt{nT}$ convergence rate then follows by Assumption 26. ■

Lemma 2.10.10 *Letting Assumption 21, 22, 23, 25, and 27 hold, we have*

$$\begin{aligned}
&\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} D_{it}\varepsilon_{1it}/\, f_{v_t}(v_{it})}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} D_{it}/f_{v_t}(v_{it})}-\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}(1-D_{it})\varepsilon_{0it}/f_{v_t}(v_{it})}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}(1-D_{it})/f_{v_t}(v_{it})} \\
= \quad & \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{D_{it}}{f_{v_t}(v_{it})\overline{\Pi}_1}\varepsilon_{1it}-\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{1-D_{it}}{f_{v_t}(v_{it})\overline{\Pi}_2}\varepsilon_{0it}+o_P\left((nT)^{-1/2}\right).
\end{aligned}$$

and $\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\left[\frac{D_{it}}{f_{v_t}(v_{it})\overline{\Pi}_1}\varepsilon_{1it}-\frac{1-D_{it}}{f_{v_t}(v_{it})\overline{\Pi}_2}\varepsilon_{0it}\right]=O_P\left((nT)^{-1/2}\right).$

Proof of Lemma 2.10.10.2 Following the same proof as in Lemma 2.10.8, we have

$$\begin{aligned}
&\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}D_{it}\varepsilon_{1it}/\,f_{v_t}(v_{it})}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}D_{it}/f_{v_t}(v_{it})}-\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}(1-D_{it})\varepsilon_{0it}/f_{v_t}(v_{it})}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}(1-D_{it})/f_{v_t}(v_{it})} \qquad (2.10.37)\\
=\;&\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\left(\frac{D_{it}}{f_{v_t}(v_{it})\overline{\Pi}_1}-1\right)\varepsilon_{1it}-\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\left(\frac{1-D_{it}}{f_{v_t}(v_{it})\overline{\Pi}_2}-1\right)\varepsilon_{0it}\\
&+\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\left(\varepsilon_{1it}-\varepsilon_{0it}-E\left(\varepsilon_{1it}-\varepsilon_{0it}|a_i,\widetilde{a}_i\right)-E\left(\varepsilon_{1it}-\varepsilon_{0it}|b_t,\widetilde{b}_t\right)\right)\\
&+\frac{1}{n}\sum_{i=1}^{n}E\left(\varepsilon_{1it}-\varepsilon_{0it}|a_i,\widetilde{a}_i\right)+\frac{1}{T}\sum_{t=1}^{T}E\left(\varepsilon_{1it}-\varepsilon_{0it}|b_t,\widetilde{b}_t\right)+o_P\left((nT)^{-1/2}\right),
\end{aligned}$$

where the first three terms are $O_p\left((nT)^{-1/2}\right)$ and last two terms are zeroby Assumption 27. So we have

$$\begin{aligned}
&\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}D_{it}Y_{1it}/\,f_{v_t}(v_{it})}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}D_{it}/f_{v_t}(v_{it})}-\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}(1-D_{it})Y_{0it}/f_{v_t}(v_{it})}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}(1-D_{it})/f_{v_t}(v_{it})}-E\left(Y_1-Y_0\right)\\
=\;&\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\left[\frac{D_{it}}{f_{v_t}(v_{it})\overline{\Pi}_1}\varepsilon_{1it}-\frac{1-D_{it}}{f_{v_t}(v_{it})\overline{\Pi}_2}\varepsilon_{0it}\right]+o_P\left((nT)^{-1/2}\right).
\end{aligned}$$

■

Lemma 2.10.11 *a_i, b_t are random vectors that satisfy Assumption 25. w_{it} are random vectors and $w_{it}\perp w_{it'}|a_i$, for $t\neq t'$, $w_{it}\perp w_{i't}|b_t$ for $i\neq i'$, $w_{it}\perp w_{i't'}$ for $i\neq i'$, $t\neq t'$. $h(a_i,b_t,w_{it})$ are a real function that the first and second moment exist, and $E\big[h(a_i,b_t,w_{it})^2\big]=$*

$o(n)$. $E[h(a_i, b_t, w_{it})] = E[h(a_{i'}, b_{t'}, w_{i't'})]$ *for any* i, t, i', t'. $T \to \infty$ *as* $n \to \infty$. *Then*

$$\frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} h(a_i, b_t, w_{it})$$

is equal to

$$E\left[h(a_i, b_t, w_{it})\right] + \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \left[E\left[h(a_i, b_t, w_{it})|a_i\right] + E\left[h(a_i, b_t, w_{it})|b_t\right] - 2E\left[h(a_i, b_t, w_{it})\right]\right] + o_p\left(\frac{1}{\sqrt{T}}\right).$$

w_{it} are heterogeneous across t, but $\mathrm{E}(h)$ are assumed the same across t. This would typically be satisfied by having $\mathrm{E}(h) = 0$ for any i, t.

Proof of Lemma 2.10.11.2 Let

$$Q = \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \left[h(a_i, b_t, w_{it}) - \mathrm{E}(h(a_i, b_t, w_{it})|a_i) - \mathrm{E}(h(a_i, b_t, w_{it})|b_t) + \mathrm{E}(h(a_i, b_t, w_{it}))\right], \tag{2.10.38}$$

To establish that $Q = o_p(\frac{1}{\sqrt{T}})$, begin with

$$\mathrm{E}[Q^2] = \frac{1}{n^2 T^2} \sum_{i=1}^{n} \sum_{t=1}^{T} \sum_{i'=1}^{n} \sum_{t'=1}^{T} \mathrm{E}\left[\left(h - \mathrm{E}(h|a_i) - \mathrm{E}(h|b_t) + \mathrm{E}(h)\right)\left(h - \mathrm{E}(h|a_{i'}) - \mathrm{E}(h|b_{t'}) + \mathrm{E}(h)\right)\right].$$

For $i \neq i', t \neq t'$, the term inside summation is zero. Now consider the case where only one index is equal to the other one, i.e., $i = i'$, $t \neq t'$. Since

$$\begin{aligned} \mathrm{E}\left[h(a_i, b_t, w_{it})h(a_i, b_{t'}, w_{it'})\right] &= \mathrm{E}\left[\mathrm{E}\left[h(a_i, b_t, w_{it})h(a_i, b_{t'}, w_{it'})|a_i\right]\right] \\ &= \mathrm{E}\left[\mathrm{E}\left[h(a_i, b_t, w_{it})|a_i\right]\mathrm{E}[h(a_i, b_{t'}, w_{it'})|a_i]\right], \end{aligned}$$

the term inside summation is zero again. So we can rewrite $\mathrm{E}[Q^2]$ as

$$\mathrm{E}[Q^2] = \frac{1}{n^2T^2}\sum_{i=1}^{n}\sum_{t=1}^{T}\mathrm{E}\left[\left(h - \mathrm{E}(h|a_i) - \mathrm{E}(h|b_t) + \mathrm{E}(h)\right)^2\right].$$

By assumption $\mathrm{E}\left(h^2\right) = o_p(n)$, so $\mathrm{E}[Q^2] = o_p\left(\frac{1}{T}\right)$, which implies $Q = o_p\left(\frac{1}{\sqrt{T}}\right)$. ■

Lemma 2.10.12 *Make the same assumptions as in Lemma 2.10.11 and Assumption 24. Further assume var*$(E\left[h(a_i, b_t, w_{it})|a_i\right]) \leq M$*, for all* i*, where* M *is a finite positive number. Then*

$$\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\left[E\left[h(a_i, b_t, w_{it})|a_i\right] + E\left[h(a_i, b_t, w_{it})|b_t\right] - 2E\left[h(a_i, b_t, w_{it})\right]\right]$$

is equal to $\frac{1}{T}\sum_{t=1}^{T}\left[E\left[h(a_i, b_t, w_{it})|b_t\right] - E\left[h(a_i, b_t, w_{it})\right]\right] + o_p\left(\frac{1}{\sqrt{T}}\right)$.

Proof of Lemma 2.10.12.2

$$\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\left[\mathrm{E}\left[h(a_i, b_t, w_{it})|a_i\right] + \mathrm{E}\left[h(a_i, b_t, w_{it})|b_t\right] - 2\mathrm{E}(h(a_i, b_t, w_{it}))\right]$$

First by assumption that $\omega_{it}|b_t$ is i.i.d across i, we know that

$$\mathrm{E}\left[h(a_i, b_t, w_{it})|b_t\right] = \mathrm{E}\left[h(a_{i'}, b_t, w_{i't})|b_t\right],$$

which gives

$$\begin{aligned} & \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\left[\mathrm{E}\left[h(a_i, b_t, w_{it})|b_t\right] - \mathrm{E}(h(a_i, b_t, w_{it}))\right] \qquad (2.10.39) \\ = \quad & \frac{1}{T}\sum_{t=1}^{T}\left[\mathrm{E}\left[h(a_i, b_t, w_{it})|b_t\right] - \mathrm{E}(h(a_i, b_t, w_{it}))\right] \end{aligned}$$

For the other part, note that

$$\begin{aligned} & \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\left[\mathrm{E}\left[h(a_i,b_t,w_{it})|a_i\right]-\mathrm{E}(h(a_i,b_t,w_{it}))\right] \\ = \quad & \frac{1}{T}\sum_{t=1}^{T}\left[\frac{1}{n}\sum_{i=1}^{n}\left[\mathrm{E}\left[h(a_i,b_t,w_{it})|a_i\right]-\mathrm{E}(h(a_i,b_t,w_{it}))\right]\right], \end{aligned}$$

where $\mathrm{E}[h(a_i,b_t,w_{it})|a_i]$ is independent across i.

$$\begin{aligned} & \mathrm{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}\left[\mathrm{E}\left[h(a_i,b_t,w_{it})|a_i\right]-\mathrm{E}(h(a_i,b_t,w_{it}))\right]\right)^2\right] \\ = \quad & \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{E}\left[\left(\left[\mathrm{E}\left[h(a_i,b_t,w_{it})|a_i\right]-\mathrm{E}(h(a_i,b_t,w_{it}))\right]\right)^2\right] \leq \frac{M}{n}, \end{aligned}$$

by Markov's inequality,

$$\frac{1}{n}\sum_{i=1}^{n}\left[\mathrm{E}\left[h(a_i,b_t,w_{it})|a_i\right]-\mathrm{E}(h(a_i,b_t,w_{it}))\right]=O_p(\frac{1}{\sqrt{n}}),$$

which gives that

$$\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\left[\mathrm{E}\left[h(a_i,b_t,w_{it})|a_i\right]-\mathrm{E}(h(a_i,b_t,w_{it}))\right]=O_p(\frac{1}{\sqrt{n}}). \tag{2.10.40}$$

The lemma then follwos from combining equation (2.10.39) and equation (2.10.40). ■

Lemma 2.10.13 *Denote* $\zeta_n=(A_{1n},B_{1n},A_{2n},B_{2n})'$, *a 4-by-1 vector, where* $A_{1n},B_{1n},A_{2n},B_{2n}$ *are random variables that evolve as n goes to infinity. Assume that* ζ_n *converge in probability to* $\overline{\zeta}=(0,\overline{B}_1,0,\overline{B}_2)'$, *where* $\overline{B}_1\neq 0,\overline{B}_2\neq 0$, *and*

$$\sqrt{n}\left[\zeta_n-\zeta\right]\xrightarrow{d}N(\mathbf{0},\mathbf{\Omega}),$$

where $\mathbf{\Omega}$ *is a positive definite matrix*

$$\Omega = \begin{pmatrix} \sigma_{A_1}^2 & \sigma_{A_1B_1} & \sigma_{A_1A_2} & \sigma_{A_1B_2} \\ . & \sigma_{B_1}^2 & \sigma_{B_1A_2} & \sigma_{B_1B_2} \\ . & . & \sigma_{A_2}^2 & \sigma_{A_2B_2} \\ . & . & . & \sigma_{B_2}^2 \end{pmatrix}.$$

Then

$$\sqrt{n}\left(\frac{A_{1n}}{B_{1n}} - \frac{A_{2n}}{B_{2n}}\right) \xrightarrow{d} N\left(0, \frac{\sigma_{A_1}^2}{\overline{B}_1^2} - \frac{2\sigma_{A_1A_2}}{\overline{B}_1\overline{B}_2} + \frac{\sigma_{A_2}^2}{\overline{B}_2^2}\right).$$

Proof.2 The Lemma follows immediately from the delta method. ■

Proof of Theorem 2.8.2.2 First we have

$$\sup_{I_{\tau it}\neq 0}\left|\widehat{f}_{v_t}(v_{it}) - f_v(v_{it})\right| = O_P\left(\log(n)\Big/\sqrt{nh}\right) = O_P\left(\log(n)\, n^{-2/5}\right).$$

Following the proof of Lemma 2.10.5, we have[7]

$$\begin{aligned} & \frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} D_{it}Y_{it}\Big/\widehat{f}_{v_t}(v_{it})}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} D_{it}\Big/\widehat{f}_{v_t}(v_{it})} - \frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}(1-D_{it})Y_{it}\Big/\widehat{f}_{v_t}(v_{it})}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}(1-D_{it})\Big/\widehat{f}_{v_t}(v_{it})} - [E(Y_1)+E(Y_0)] \\ = & \frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\Lambda_{1it}}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\Pi_{1it}} - \frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\Lambda_{2it}}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\Pi_{2it}} + o_P\left(\frac{1}{\sqrt{n}}\right). \end{aligned}$$

[7] Note that the residual here is $o_P\left(\frac{1}{\sqrt{n}}\right)$. We do not need this to be $o_P\left(\frac{1}{\sqrt{nT}}\right)$ due to the slower convergence of our estimator.

Applying Lemma 2.10.12 on this expression, it is equivalent to

$$\frac{\frac{1}{T}\sum_{t=1}^{T} E\left[\Lambda_{1it}|\, b_t, \widetilde{b}_t\right]}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\Pi_{1it}} - \frac{\frac{1}{T}\sum_{t=1}^{T} E\left[\Lambda_{2it}|\, b_t, \widetilde{b}_t\right]}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\Pi_{2it}} + o_p\left(\frac{1}{\sqrt{T}}\right).$$

Applying Lemma 2.10.13 to this expression, we have

$$\begin{aligned}
& \frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{D_{it}Y_{it}}{\widehat{f}_{v_t}(v_{it})}}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{D_{it}}{\widehat{f}_{v_t}(v_{it})}} - \frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{(1-D_{it})Y_{it}}{\widehat{f}_{v_t}(v_{it})}}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{(1-D_{it})}{\widehat{f}_{v_t}(v_{it})}} - E(\widetilde{a}_i + \widetilde{b}_t + Y_1) + E(\widetilde{a}_i + \widetilde{b}_t + Y_0) \\
= & \frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\Lambda_{1it}}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\Pi_{1it}} - \frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\Lambda_{2it}}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\Pi_{2it}} + o_p\left(\frac{1}{\sqrt{n}}\right) \\
= & \frac{\frac{1}{T}\sum_{t=1}^{T} E\left[\Lambda_{1it}|\, b_t, \widetilde{b}_t\right]}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\Pi_{1it}} - \frac{\frac{1}{T}\sum_{t=1}^{T} E\left[\Lambda_{2it}|\, b_t, \widetilde{b}_t\right]}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\Pi_{2it}} + o_p\left(\frac{1}{\sqrt{T}}\right),
\end{aligned}$$

which then gives the conclusion by applying Lemma 2.10.13. ■

2.10.2 Proof of Theorem 2.5.1, 2.5.3 and 2.5.5

Lemma 2.10.14 *$M - v_n^{(1)} \propto n^{-1}$ in probability.*

Proof.2 Let $\{a_n\}_{n=1}^{\infty}$ be any series that $a_n \to \infty$ and $a_n = o(n)$. Let $\underline{c}_v = \inf_{v\in \mathrm{supp}(V)} f_v(V)$. Then

$$P\left(v_n^{(1)} < M - \frac{a_n}{n}\right) \leq \left(1 - \underline{c}_v \frac{a_n}{n}\right)^n = \left(\left(1 - \underline{c}_v \frac{a_n}{n}\right)^{\frac{n}{\underline{c}_v a_n}}\right)^{\underline{c}_v a_n} = \left(e\left(1 + o\left(1\right)\right)\right)^{-\underline{c}_v a_n} \to 0,$$

where the second equality holds by the fact that $\lim_{x\to 0}(1-x)^{\frac{1}{x}} = e^{-1}$. So we have $M - v_n^{(1)} = O_P\left(n^{-1}\right)$. Let $\overline{c}_v = \sup_{v\in \text{supp}(V)} f_v(V)$. On the other hand, if $a_n \to 0$, then

$$P\left(v_n^{(1)} < M - \frac{a_n}{n}\right) \geq \left(1 - \overline{c}_v \frac{a_n}{n}\right)^n = \left(e\left(1 + o\left(1\right)\right)\right)^{-\underline{c}_v a_n} \to 1.$$

So we have in probability $M - v_n^{(1)} \propto n^{-1}$. ■

Proof of Theorem 2.5.1.2 The proof of this theorem is standard. We define the components of the bias term and variance term from the estimates by $\mathbf{B}_h$ and $\mathbf{V}_h$ respectively:

$$\begin{aligned}
\mathbf{B}_h\left(\widehat{M}\right) &\equiv \frac{1}{n-1}\sum_{i=1}^{n-1} K_h\left(V_i - \widehat{M}\right) \begin{pmatrix} 1 \\ \left(V_i - \widehat{M}\right)/h \end{pmatrix} \left[G_D\left(V_i\right) - G_D\left(\widehat{M}\right) - G_D'\left(\widehat{M}\right)\left(V_i - V\right)\right], \\
\mathbf{V}_h\left(\widehat{M}\right) &\equiv \frac{1}{n-1}\sum_{i=1}^{n-1} K_h\left(V_i - \widehat{M}\right) \begin{pmatrix} 1 \\ \left(V_i - \widehat{M}\right)/h \end{pmatrix} \left[I\left(D_i = 1\right) - G_D\left(V_i\right)\right].
\end{aligned}$$

Then $\widehat{G}_D\left(\widehat{M}\right) = G_D\left(\widehat{M}\right) + e_1^T\left[\mathbf{S}_h\left(\widehat{M}\right)\right]^{-1}\left(\mathbf{B}_h\left(\widehat{M}\right) + \mathbf{V}_h\left(\widehat{M}\right)\right)$.

One can then show that

$$\begin{aligned}
\mathbf{S}_h\left(\widehat{M}\right) &\xrightarrow{P} \overline{\mathbf{S}}, \\
\mathbf{B}_h\left(\widehat{M}\right) &= h^2 \begin{pmatrix} S_{2,-} \\ S_{3,-} \end{pmatrix} G_D''\left(M\right) f_v\left(M\right) + o_P\left(h^2\right),
\end{aligned}$$

and

$$\begin{aligned}
E\left[\mathbf{V}_h\left(\widehat{M}\right)\right] &= 0 \\
E\left[\mathbf{V}_h\left(\widehat{M}\right)^2\right] &\equiv \mathbf{Q} G_D\left(M\right)\left(1 - G_D\left(M\right)\right) f_v\left(M\right) + o\left(1\right).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\text{bias}\left(\widehat{G}_D\left(\widehat{M}\right)\right) &= e_1^T \begin{pmatrix} S_{2,-} \\ S_{3,-} \end{pmatrix} G_D''(M) f_v(M) h^2 + o_P\left(h^2\right), \\
\text{var}\left(\widehat{G}_D\left(\widehat{M}\right)\right) &= \frac{1}{nh} e_1^T \overline{\mathbf{S}}^{-1} \mathbf{Q} \overline{\mathbf{S}}^{-1} e_1 G_D(M)(1 - G_D(M)) f_v(M) + o\left(\frac{1}{nh}\right),
\end{aligned}$$

where the leading terms in the bias and variance are $\mathbb{B}_h$ and $\sigma^2(M)$ respectively.

By Lemma 2.10.14, $G_D\left(\widehat{M}\right) - G_D(M) = O_P\left(n^{-1}\right)$, thus

$$\sqrt{nh}\left(\widehat{G}_D\left(\widehat{M}\right) - G_D(M) - \mathbb{B}_h\right) \xrightarrow{d} N\left(0, \sigma^2(M)\right),$$

which is the conclusion.

Since $\text{MSE}\left(\widehat{G}_D\left(\widehat{M}\right)\right) = \left[\text{bias}\left(\widehat{G}_D\left(\widehat{M}\right)\right)\right]^2 + \text{var}\left(\widehat{G}_D\left(\widehat{M}\right)\right)$, to minimize mean squared error we can get h_{opt} as

$$h_{\text{opt}} = n^{-1/5}\left[\left(e_1^T \overline{\mathbf{S}}^{-1} \mathbf{Q} \overline{\mathbf{S}}^{-1} e_1 G_D(M)(1 - G_D(M)) f_v(M)\right) \Big/ \left(e_1^T \overline{\mathbf{S}}^{-1} \begin{pmatrix} S_{2,-} \\ S_{3,-} \end{pmatrix} G_{D,-}''(M) f_v(M)\right)^2\right]^{-1/5}$$

■

Proof of Theorem 2.5.3.2 Most proof of this theorem is standard, except that $\widehat{G}_D\left(\widehat{M}\right)$ converges at the $\sqrt{n}$ rate, while in the typical case the convergence rate would be $\sqrt{nh}$. The intuition for this result is that in $\mathbf{V}_h\left(\widehat{M}\right)$ $E\left[\left(I(D_i = 1) - G_D(V_i)\right)^2 \middle| V_i = M\right] = 0$, and $E\left[\left(I(D_i = 1) - G_D(V_i)\right)^2 \middle| V_i = M - h\right] \propto h$, under the large support assumption. This is because we put the most weight on the observations around M within ch during estimation, for some $c > 0$. The variance for those observations is of the order h, resulting in the faster

rate of convergence.

Comparing this to the proof of Theorem 2.5.1, the difference is that

$$E\left[\mathbf{V}_h\left(\widehat{M}\right)^2\right]=\frac{1}{n}\mathbf{Q}G_D'\left(M\right)f_v\left(M\right)+\frac{\widehat{M}-M}{nh}\mathbf{Q}G_D'\left(M\right)f_v\left(M\right)+o_P\left(\frac{1}{n}\right)+o_P\left(\frac{\widehat{M}-M}{nh}\right),$$

where the leading term in $E\left[\mathbf{V}_h\left(\widehat{M}\right)^2\right]$ in Theorem 2.5.1 becomes zero here.

Therefore,

$$\begin{aligned}\text{bias}\left(\widehat{G}_D\left(\widehat{M}\right)\right) &= h^2e_1^T\begin{pmatrix}S_{2,-}\\ \\ S_{3,-}\end{pmatrix}G_D''\left(M\right)f_v\left(M\right)+o_P\left(h^2\right),\\ \text{var}\left(\widehat{G}_D\left(\widehat{M}\right)\right) &= \frac{1}{n}e_1^T\mathbf{Q}e_1G_D'\left(M\right)f_v\left(M\right)+\frac{\widehat{M}-M}{nh}e_1^T\mathbf{Q}e_1G_D'\left(M\right)f_v\left(M\right)+o_P\left(\frac{1}{n}\right)+o_P\left(\frac{\widehat{M}-M}{nh}\right),\end{aligned}$$

here the leading terms in bias and variance are $\mathbb{B}_h$ and $\widetilde{\sigma}^2\left(M\right)$, respectively.

Since $\text{MSE}\left(\widehat{G}_D\left(\widehat{M}\right)\right)=\left[\text{bias}\left(\widehat{G}_D\left(\widehat{M}\right)\right)\right]^2+\text{var}\left(\widehat{G}_D\left(\widehat{M}\right)\right)$, minimize means squared error we can get h_{opt} as

$$h_{\text{opt}}=\left(\frac{\widehat{M}-M}{n}\right)^{1/5}e_1^T\mathbf{Q}e_1G_D'\left(M\right)\Bigg/\left[\left(e_1^T\begin{pmatrix}S_{2,-}\\ \\ S_{3,-}\end{pmatrix}G_D''\left(M\right)\right)^2f_v\left(M\right)\right].$$

By Lemma 2.10.14 $\widehat{M}-M\propto n^{-1}$ in probability, so $h_{\text{opt}}\propto n^{-2/5}$. Therefore, the bias term is asymptotically negligible and we have $\sqrt{n}\left(\widehat{G}_D\left(\widehat{M}\right)-G_D\left(\widehat{M}\right)\right)\xrightarrow{d}N\left(0,\sigma^2\left(M\right)\right)$. By Lemma 2.10.14 again, $G_D\left(\widehat{M}\right)-G_D\left(M\right)=O_P\left(n^{-1}\right)$, and thus

$$\sqrt{n}\left(\widehat{G}_D\left(\widehat{M}\right)-G_D\left(\widehat{M}\right)\right)\xrightarrow{d}N\left(0,\sigma^2\left(M\right)\right),$$

which is the conclusion. ■

Proof of Theorem 2.5.5.2 By Assumption 29, $E(Y_0|X) = E(Y_0|X, V \leq -\gamma_n(X))$. The first part of the theorem follows from

$$\begin{aligned} & \lim_{n\to\infty} E(I(D=0)Y_0|X, V \leq -\gamma_n(X)) - E(Y_0|X, V \leq -\gamma_n(X)) \\ = & \lim_{n\to\infty} E[(I(D=0)-1)Y_0|X, V \leq -\gamma_n(X)] = 0, \end{aligned}$$

where the last equality holds by $\lim_{n\to\infty} E(D|X, V \leq -\gamma_n(X)) = 0$. This generates the expression for $E(Y_0|X)$, and the expression for $E(Y_2|X)$ is obtained in the same way. The expression for $E(Y_1|X)$ follows immediately from Theorem 2.3.2. ■

2.10.3 Proof of Theorem 2.8.1 and 2.8.8

Proof of Theorem 2.8.1.2 First, the following is identified:

$$\begin{aligned} & E(D|V=v, Z=z, X=x) \\ = & F_{U|X}(\alpha_1(x) - \varsigma(v) - \varpi(x,z)|x) - F_{U|X}(\alpha_0(x) - \varsigma(v) - \varpi(x,z)|x), \\ & \frac{\partial E(D|V=v, Z=z, X=x)}{\partial v} \\ = & -\left[f_{U|X}(\alpha_1(x) - \varsigma(v) - \varpi(x,z)|x) - f_{U|X}(\alpha_0(x) - \varsigma(v) - \varpi(x,z)|x)\right]\frac{d\varsigma(v)}{dv}, \\ & \frac{\partial E(D|V=v, Z=z, X=x)}{\partial z} \\ = & -\left[f_{U|X}(\alpha_1(x) - \varsigma(v) - \varpi(x,z)|x) - f_{U|X}(\alpha_0(x) - \varsigma(v) - \varpi(x,z)|x)\right]\frac{\partial \varpi(x,z)}{\partial z}. \end{aligned}$$

$\frac{d\varsigma(v)}{dv} \Big/ \frac{\partial \varpi(x,z)}{\partial z}$ is identified by

$$\frac{d\varsigma\left(v\right)}{dv} \Big/ \frac{\partial \varpi\left(x,z\right)}{\partial z} = \frac{\partial E\left(D|V=v,Z=z,X=x\right)}{\partial v} \Big/ \frac{\partial E\left(D|V=v,Z=z,X=x\right)}{\partial z}. \tag{2.10.41}$$

Then fix $V=0$, by $\varsigma'(0)=1$, and $\frac{\partial \varpi(x,z)}{\partial z}$ is identified by varying (X,Z). Fix X,Z at some point, and then by knowing $\frac{\partial \varpi(x,z)}{\partial z}$, $\varsigma'(v)$ is identified. Finally, $\varsigma(V)$ is identified by

$$\varsigma(v) = \varsigma(0) + \int_0^v \varsigma'(s)\, ds.$$

■

Proof of Theorem 2.8.8.2 The proof here is very similar to the proof of Theorem 2.3.4.

Start by looking at

$$\begin{aligned}
& E\left(\frac{D_{it}Y_{it}}{f_{v_t}(V_{it}|X_{it},V_{it-1})} \middle| U_{it}, a_i, b_t, X_{it}, D_{it-1}, V_{it-1} \right) \\
= \; & E\left[E\left(\frac{D_{it}\left(\widetilde{a}_i + \widetilde{b}_t + Y_{1it} + g\left(Y_{it-1}\right)\right)}{f_{v_t}(V_{it}|X_{it},V_{it-1})} \mid V_{it}, U_{it}, a_i, b_t, X_{it}, D_{it-1}, V_{it-1} \right) \mid U_{it}, a_i, b_t, X_{it}, D_{it-1}, V_{it-1} \right] \\
= \; & E\left[\frac{I\left(\alpha_0(X_{it}) \leq a_i + b_t + V_{it} + \vartheta\left(D_{it-1}\right) + U_{it} \leq \alpha_1(X_{it})\right)}{f_{v_t}(V_{it}|X_{it},V_{it-1})} \right. \\
& \left. E\left(\widetilde{a}_i + \widetilde{b}_t + Y_{1it} + g\left(Y_{it-1}\right) \mid V_{it}, U_{it}, a_i, b_t, X_{it}, D_{it-1}, V_{it-1}\right) \mid U_{it}, a_i, b_t, X_{it}, D_{it-1}, V_{it-1} \right] \\
= \; & \int_{supp(V_{it}|U_{it},a_i,b_t,X_{it},D_{it-1},V_{it-1})} \frac{I\left(\alpha_0(X_{it}) \leq a_i + b_t + V_{it} + \vartheta\left(D_{it-1}\right) + U_{it} \leq \alpha_1(X_{it})\right)}{f_{v_t}(v_{it}|X_{it},V_{it-1})} \\
& E\left(\widetilde{a}_i + \widetilde{b}_t + Y_{1it} + g\left(Y_{it-1}\right) \mid U_{it}, a_i, b_t, X_{it}, D_{it-1}, V_{it-1}\right) f_{v_t}\left(v_{it} \mid U_{it}, a_i, b_t, X_{it}, D_{it-1}, V_{it-1}\right) dv_{it} \\
= \; & \int_{\alpha_0(X_{it})-a_i-b_t-U_{it}-\vartheta(D_{it-1})}^{\alpha_1(X_{it})-a_i-b_t-U_{it}-\vartheta(D_{it-1})} E\left(\widetilde{a}_i + \widetilde{b}_t + Y_{1it} + g\left(Y_{it-1}\right) \mid U_{it}, a_i, b_t, X_{it}, D_{it-1}, V_{it-1}\right) dv_{it} \\
= \; & E\left(\widetilde{a}_i + \widetilde{b}_t + Y_{1it} + g\left(Y_{it-1}\right) \mid U_{it}, a_i, b_t, X_{it}, D_{it-1}, V_{it-1}\right) \int_{\alpha_0(X_{it})-a_i-b_t-U_{it}-\vartheta(D_{it-1})}^{\alpha_1(X_{it})-a_i-b_t-U_{it}-\vartheta(D_{it-1})} 1 dv_{it} \\
= \; & E\left(\widetilde{a}_i + \widetilde{b}_t + Y_{1it} + g\left(Y_{it-1}\right) \mid U_{it}, a_i, b_t, X_{it}, D_{it-1}, V_{it-1}\right) \left[\alpha_1(X_{it}) - \alpha_0(X_{it})\right]
\end{aligned}$$

and therefore

$$
\begin{aligned}
& E\left[D_{it}Y_{it}/f_{v_t}(V_{it}|X_{it},V_{it-1})|X_{it}\right] \\
= \; & E\left[E\left(\widetilde{a}_i+\widetilde{b}_t+Y_{1it}+g\left(Y_{it-1}\right)\mid U_{it},a_i,b_t,X_{it},D_{it-1},V_{it-1}\right)\left[\alpha_1(X_{it})-\alpha_0(X_{it})\right]|X_{it}\right] \\
= \; & E\left(Y_{1it}+\widetilde{a}_i+\widetilde{b}_t+g\left(Y_{it-1}\right)\Big| X_{it}\right)\left[\alpha_1(X_{it})-\alpha_0(X_{it})\right].
\end{aligned}
$$

Given the above result, the rest of the proof follows from the same logic as the proof for Theorem 2.3.2. ■

Bibliography

[1] Powell, J. L., J. H. Stock, and T. M. Stoker (1989), "Semiparametric Estimation of Index Coeffients," Econometrica, 57, 1403-1430.

[2] Silverman, B. W. (1978), "Weak and Strong Uniform Consistency of the Kernel Estimate of a Density Function and its Derivatives," Annals of Statistics, 6, 177-184.

2.10.4 Additional Tables

Table 4: Monte Carlo results matching the empirical data

	MEAN(−3.9)	SD	LQ	MED	UQ	RMSE	MAE	MDAE	%2SE
Panel A: Symmetric setting with normal errors									
Trim-ATE	−3.90	0.43	−4.19	−3.90	−3.61	0.43	0.34	0.00	1.00
No-Trim-ATE	−3.90	1.22	−4.67	−3.92	−3.12	1.22	0.95	0.02	1.00
Naive-ATE	−3.90	0.32	−4.11	−3.90	−3.68	0.32	0.25	0.00	1.00
ML-ATE	−3.90	0.30	−4.10	−3.90	−3.70	0.30	0.24	0.00	1.00
Panel B: Symmetric setting with uniform errors									
Trim-ATE	−3.90	0.38	−4.16	−3.90	−3.64	0.38	0.31	0.00	1.00
No-Trim-ATE	−3.90	0.38	−4.16	−3.90	−3.64	0.38	0.31	0.00	1.00
Naive-ATE	−3.90	0.38	−4.16	−3.90	−3.65	0.38	0.30	0.00	1.00
ML-ATE	−3.91	0.38	−4.17	−3.90	−3.65	0.38	0.30	0.00	1.00
Panel C: Asymmetric setting with normal errors									
Trim-ATE	−3.21	0.51	−3.55	−3.21	−2.87	0.86	0.73	0.69	0.95
No-Trim-ATE	−3.65	1.33	−4.50	−3.65	−2.81	1.35	1.06	0.25	0.77
Naive-ATE	−1.99	0.34	−2.21	−2.00	−1.77	1.94	1.91	1.90	0.15
ML-ATE	−1.98	0.35	−2.22	−1.98	−1.75	1.95	1.92	1.92	0.15
Panel D: Asymmetric setting with uniform errors									
Trim-ATE	−3.45	0.48	−3.77	−3.45	−3.12	0.66	0.54	0.45	0.99
No-Trim-ATE	−3.76	1.08	−4.47	−3.76	−3.06	1.09	0.86	0.14	0.85
Naive-ATE	−1.84	0.37	−2.08	−1.84	−1.59	2.10	2.06	2.06	0.09
ML-ATE	−2.07	0.39	−2.34	−2.07	−1.81	1.87	1.83	1.83	0.25

Note: True E(Y_1)−E(Y_0) = −3.9. Parameters set (θ_0, θ_1, θ_{01}, θ_{02}, θ_{11}, θ_{12}, θ_2) for the four MC in order are as follows: (6.94 3.04 5.64 8.44 6.71 4.87 1.06), (6.97 3.07 23.67 −24.30 22.62 25.72 1.07), (6.67 2.77 6.57 −2.91 4.51 −5.43 0.43), (7.41 3.51 8.43 −4.27 5.47 −1.47 0.55). Trim-ATE and No-Trim-ATE are our proposed estimator with and without trimming (2%) respectively. Naive-ATE is an estimate for $E(Y_1|T = 1) - E(Y_0|T = 0)$. ML-ATE is Heckman's selection MLE. All statistics are for the simulation estimates. MEAN = mean. SD = standard errors. LQ = 25% quantile (lower). MED = 50% quantile (median). UQ = 75% quantile (upper). RMSE = root mean square errors. MAE = mean absolute errors. MDAE = median absolute errors. %2SE = percentage of simulations in which the true coefficient was within two estimated standard errors of the estimated coefficient.

Table 5: Robust check: Monte Carlo with normal errors

	Quadratic			Step		
	MEAN (≈ -3.9)	SD	RMSE	MEAN (-3.9)	SD	RMSE
Panel A: $\kappa_1 = 0.02$, Noise Ratio = 0.19						
Trim-ATE	-4.23	0.46	0.49	-3.19	0.41	0.82
No-Trim-ATE	-7.79	1.57	4.20	-3.31	1.04	1.20
Naive-ATE	-3.75	0.38	0.39	-3.14	0.34	0.83
ML-ATE	-3.67	0.73	0.76	-3.10	0.66	1.04
Control Function	-3.74	0.24	0.31	-1.38	0.20	2.52
Panel B: $\kappa_1 = 0.03$, Noise Ratio = 0.28						
Trim-ATE	-4.08	0.42	0.42	-2.85	0.42	1.11
No-Trim-ATE	-7.68	1.61	4.11	-2.96	1.11	1.46
Naive-ATE	-3.60	0.37	0.47	-2.79	0.34	1.16
ML-ATE	-3.54	0.74	0.82	-2.74	0.64	1.33
Control Function	-3.59	0.23	0.41	-1.33	0.21	2.58
Panel C: $\kappa_1 = 0.04$, Noise Ratio = 0.36						
Trim-ATE	-3.93	0.48	0.48	-2.55	0.42	1.41
No-Trim-ATE	-7.63	1.64	4.07	-2.66	1.09	1.65
Naive-ATE	-3.40	0.38	0.62	-2.45	0.34	1.49
ML-ATE	-3.33	0.66	0.87	-2.42	0.59	1.60
Control Function	-3.40	0.26	0.59	-1.26	0.20	2.62

Note: True mean value is -3.9. Noise ratio is defined as the ratio of standard deviation of c_e to the standard deviation of $c*$. The first three and last three columns are the results when the true response forms are quadratic and step function respectively. Five different estimators are reported here. Trim-ATE and No-Trim-ATE are our proposed estimator with and without trimming (2%) respectively. Naive-ATE is an estimate for $E(Y_1|T=1) - E(Y_0|T=0)$. ML-ATE is Heckman's selection MLE. Control function approach is defined as in the paper. MEAN = mean. SD = standard errors. RMSE = root mean square errors.

Table 6: Robust check: Monte Carlo with uniform errors

	Quadratic			Step		
	MEAN (≈ -3.9)	SD	RMSE	MEAN (-3.9)	SD	RMSE
Panel A: $\kappa_2 = 0.06$, Noise Ratio = 0.17						
Trim-ATE	-3.86	0.36	0.36	-3.23	0.34	0.76
No-Trim-ATE	-3.96	0.36	0.36	-3.23	0.34	0.75
Naive-ATE	-3.79	0.34	0.35	-3.24	0.34	0.74
ML-ATE	-3.54	1.76	1.79	-3.23	0.51	0.84
Control Function	-3.71	0.25	0.31	-1.87	0.23	2.04
Panel B: $\kappa_2 = 0.07$, Noise Ratio = 0.19						
Trim-ATE	-3.83	0.35	0.35	-3.13	0.34	0.84
No-Trim-ATE	-3.92	0.35	0.35	-3.14	0.33	0.83
Naive-ATE	-3.76	0.35	0.37	-3.13	0.34	0.84
ML-ATE	-3.46	1.87	1.92	-3.10	0.55	0.97
Control Function	-3.65	0.25	0.35	-1.84	0.23	2.07
Panel C: $\kappa_2 = 0.08$, Noise Ratio = 0.22						
Trim-ATE	-3.79	0.36	0.37	-3.04	0.33	0.91
No-Trim-ATE	-3.88	0.36	0.36	-3.05	0.33	0.91
Naive-ATE	-3.70	0.35	0.39	-3.02	0.33	0.94
ML-ATE	-3.40	1.85	1.91	-3.02	0.53	1.03
Control Function	-3.59	0.25	0.40	-1.82	0.23	2.10

Note: True mean value is -3.9. Noise ratio is defined as the ratio of standard deviation of c_e to the standard deviation of $c*$. The first three and last three columns are the results when the true response forms are quadratic and step function respectively. Five different estimators are reported here. Trim-ATE and No-Trim-ATE are our proposed estimator with and without trimming (2%) respectively. Naive-ATE is an estimate for $E(Y_1|T=1) - E(Y_0|T=0)$. ML-ATE is Heckman's selection MLE. Control function approach is defined as in the paper. MEAN = mean. SD = standard errors. RMSE = root mean square errors.

Chapter 3

Binary choice model with interactive effects

With Qiankun Zhou

3.1 Introduction

Nowadays, econometric analysis of models with interactive effects or cross sectional dependence has gained lots of attention both theoretically and empirically. The interactive effects, or cross sectional dependence, is used to capture the unobserved individual and time-specific effects. Compared to models without interactive effects, the model with interactive effects provide a more reliable estimator (for example, see Bai (2003, 2009a), Bai and Ng (2002, 2008)). Moreover, taking interactive effects into account would also reduce the heterogeneity of the model and thus eliminate the source of bias in panel data models (Hsiao (2014)). A number of different approaches have been advanced for dealing with models with interactive effects, among which Pesaran (2006) proposes the so called common correlated effects

(CCE) estimator which can be computed by least squares in augmented regressions with cross-sectional averages of the dependent variable and the individual-specific regressors, and Bai (2009a) investigates identification and estimation of panel data model with interactive effects through the principal component approach. Other approaches can be found in Bai and Serene (2008) and the reference therein.

For the above approaches of dealing with models with interactive effects, there are several issues needed to addressed. On the one hand, these approaches usually assume the model is linear, and it would be problematic if they are applied to nonlinear model (for example, binary choice model), on the other, these approaches usually assume large N and large T when deriving the limiting behavoir of the estimator, but it's rare the case that econometricans have enough time period data in microeconometrics where the time periods are usually small. As a result, it would be necessary to extend the previous works on dealing with interactive effects to the case where the model is nonlinear and the time periods are small or fixed.

In this paper, we consider the estimation of binary choice model with interactive effects when the number of cross-section units N is large and the number of time periods T is fixed. The various applications of binary choice model has its root in microeconometrics where economists usually have interest to investigate the plausibility of some specific policies or programs. In most cases, the outcome of the policies and programs can be normalized as a zero-one variable which suits the setup of binary choice model. Hsiao (2014) provides a general application of binary choice model. Also, for empirical analysis in microeconometrics, there are typically large amount of cross sectional individuals such as surveys from households, but the length of survey is always small or fixed, for example, the PSID study

contains thousands of individuals in the past 50 years. Consequently, we only consider the case when T is a fixed number. When T is large, i.e., going infinity as sample size increases, our results could be extended without much difficulty.

Unlike the usual methods of dealing with interactive effect as in Bai (2009a) and Pesaran (2006), our approach relies on projection methods. Especially, we use the projection method of Mundlak (1978) to control the cross sectional dependence, this approach has been recently considered by Bai (2009b). The projection method is widely used in econometrics to model the unobservable effects with the observables of the model, for example, Hayakawa (2012) and Semykina and Wooldridge (2010) and the related reference. This paper also applies the so called special regressor method proposed by Lewbel (2000a) and Honore and Lewbel (2002)[1], which transforms the nonlinear model into a linear one. Upon transformation, we use the usual partition regression method to obtain the estimator of parameters of interest. Obviously, our estimator has the advantage of computational simplicity compared to the estimator recently proposed by Fernandez-Val and Weidner (2012), where there is no closed form for the estimators and nonlinear optimization is needed for calculation.

We also develop asymptotic theory for the special regressor estimator of large N and fixed T. Monte Carlo simulation shows that the special regressor method outperforms the MLE in the presence of interactive effects. Finally, we consider the application of our approach to the women's laborforce participation. Compared to the existing researches on the women's laborforce participation, our approaches suggest that husbands' income have significant negative effects on the women's laborforce participation rather than nonsignificant effects. Our finding is intuitive since it's normal that women are less likely to work when

[1]For recent works of special regressor method, refer to Dong and Lewbel (2012), Lewbel (2012) and Lewbel et al (2012).

husbands' income is high.

The rest of the paper is organized as follows: Section 2 introduces the models, assumptions and motivational examples. Section 3 provides the estimation procedure as well as the asymptotic analysis. Section 4 reports the results of the Monte Carlo simulation. An empirical application to women's laborforce participation is provided in Section 5. Section 6 concludes by identifying important areas for extensions and further developments. All proofs are given in the appendix.

3.2 Model

3.2.1 Setup

We begin by considering the following discrete choice model with interactive effects

$$y_{it}^{*} = v_{it} + \delta_t + x_{it}'\beta + u_{it}, \quad t = 1, \ldots, T; i = 1, \ldots, N \tag{3.2.1}$$

$$u_{it} = \lambda_i' f_t + \varepsilon_{it} \tag{3.2.2}$$

$$y_{it} = 1\{y_{it}^{*} > 0\} \tag{3.2.3}$$

where y_{it} be the observation on the i-th cross-section unit at time t, δ_t is time effect, and x_{it} is a $k \times 1$ vector of observed individual-specific regressors on the i-th cross-section unit at time t, λ_i and f_t are each $r \times 1$ and both are unobservable, and ε_{it} is the error term. $1\{A\}$ is the indicator function and takes value one if condition A is satisfied and zero otherwise. The number of factors r is fixed. Moreover, we assume v_{it} is a special regressor, which satisfies the following conditions: (i) v_{it} is a continuous random variable; (ii) v_{it} is independent of δ_t and u_{it} conditional on x_{it}; (iii) v_{it} has a relatively large support. These conditions will

be elaborated more in the following sections.

Example 3.2.1 *The model considered above has its roots in economics, especially in microeconometrics. As pointed by Bai (2009b), in microeconometrics, for example, if we want to conduct a survey to study whether or not the workers will accept the job offer based on the salaries. In this kind of survey, we can use an indicator of 1 and 0 to denote the final decision, and the observed wage is a function of observable variables* (x_{it}) *and unobserved innate ability* (λ_i)*. The innate ability is potentially correlated with the observed individual characteristics such as education. It is also assumed that the innate ability is priced at each period such that its effect on wage is time varying which can be captured by* f_t*. The consequence for this motivation is a factor analytic error structure that is correlated with the regressors. Moreover, multiple factors could also be considered to allow wages to be affected by other unobservable individual traits such as dedication and perseverance.*

In this paper, we will focus on the situation in which the number of cross-section units (N) is large and the number of time periods T is fixed. For this approach, because T is small, it is desirable to treat f_t as parameters instead of treating λ_i as parameters where both f_t and λ_i are unobservable individual and time effects.

The primary interest of the present paper is the correlation between λ_i and the regressors which is motivated from the above example. As a result, projection method used for modeling unobservables with observables (for recent application of projection method, refer to Bai (2009b), Hayakawa (2012), Semykina and Wooldridge (2010)) can be applied here. Following Chamberlain (1982) as well as Bai (2009b), we assume that

$$E\left(\lambda_i \left| x_{i1}, x_{i2}, \ldots, x_{iT} \right.\right) = \lambda + \sum_{s=1}^{T} \psi_s x_{is} \tag{3.2.4}$$

where λ is a $r \times 1$ vector and ψ_s is an $r \times k$ matrix ($s \geq 1$). Equivalently, we can view the above as a linear projection, and we can observe that the problem of the above projection is that there are too many parameters to estimate. Instead, we can consider a restricted version of projection (Mundlak, 1978) as follows

$$E\left(\lambda_i \left| \bar{x}_i \right.\right) = \lambda + \psi \bar{x}_i \tag{3.2.5}$$

with $\bar{x}_i = \frac{1}{T}\sum_{t=1}^{T} x_{it}$ and ψ is an $r \times k$ matrix. And we can write the above projection as following model

$$\lambda_i = \lambda + \psi \bar{x}_i + \eta_i$$

where, by definition, we have $E\left(\eta_i \left| x_{i1}, x_{i2}, \ldots, x_{iT} \right.\right) = 0$.

Using Mundlak's projection, model (3.2.1)-(3.2.2) can be rewritten as

$$y_{it}^{*} = v_{it} + (\delta_t + \lambda f_t) + x_{it}'\beta + \bar{x}_i'\psi' f_t + f_t'\eta_i + \varepsilon_{it}$$

and we can still use δ_t for $\delta_t + \lambda f_t$ for simplicity of notations, i.e.,

$$y_{it}^{*} = v_{it} + \delta_t + x_{it}'\beta + \bar{x}'\psi' f_t + f_t'\eta_i + \varepsilon_{it} \tag{3.2.6}$$

Substitute equation (3.2.6) into (3.2.3) we have

$$\begin{aligned} y_{it}^{*} &= v_{it} + \delta_t + x_{it}'\beta + \bar{x}_i'\psi' f_t + f_t'\eta_i + \varepsilon_{it} \\ y_{it} &= 1\left\{y_{it}^{*} > 0\right\} \quad t = 1, \ldots, T; i = 1, \ldots, N \end{aligned} \tag{3.2.7}$$

Remark 3.2.2 For the Mundlak's projection method of equation (3.2.6), it's closely related to the augmented regression method proposed by Pesaran (2006), where Pesaran suggests to approximate f_t by observable proxies (δ_t, x_{it}).

The parameter of interest is β, not $f_1, \ldots, f_T$ and ψ. In order to estimate β, we need to impose several assumption on model (3.2.7), and we follow Bai (2009b)'s way to do so. To simplify notation, let $e_{it} \equiv f_t'\eta_i + \varepsilon_{it}$ and denote the conditional distribution of e_{it} conditional on $x_{it}, \overline{x}_i$ as $F_{e_{it}}(e_{it}\,|x_{it}\,)$ with the support $\Omega_{e_{it}}$.

Assumption 1: $(\boldsymbol{x}_i, \boldsymbol{\eta}_i, \boldsymbol{\varepsilon}_i)$ are *iid* over i where $\boldsymbol{x}_i \equiv (x_{i1}, x_{i2}, \ldots, x_{iT})'$ and $\boldsymbol{\varepsilon}_i \equiv (\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{iT})'$. The rank of $E(\boldsymbol{x}_i'\boldsymbol{x}_i) = k$, i.e., $E(\boldsymbol{x}_i'\boldsymbol{x}_i)$ is of full rank.

Assumption 2: $E\left(e_{it}|x_{it}, \overline{x}_i\right) \equiv 0$.

For the special regressor, v_{it}, we shall impose the following assumptions about its support and distribution, all of these assumption are standard in the literature for special regressors, for instance, Lewbel (2000a), Honoré and Lewbel (2002), Liang (2011), etc,. More specifically, we assume that

Assumption (S1): The conditional distribution of v_{it} given x_{it} has a continuous conditional density function $f_t(v_{it}\,|x_{it}\,)$ with respect to Lebesgue measure on the real line. The support of v_{it} conditional on x_{it},is $[L_t, K_t]$ where $-\infty \leq L_t < 0 < K_t \leq \infty$, and $\inf_{v_{it}\in[L_t,K_t]} f_t(v_{it}\,|x_{it}\,) > 0$.

Assumption (S2): $\delta_t, \eta_i, \varepsilon_{it} \perp v_{it}|\, x_{it}, \overline{x}_i$ and $f_t\left(v_{it}|\, x_{it}, \overline{x}_i\right) = f_t\left(v_{it}|\, x_{it}\right)$.

Assumption (S3): The support of $s_{it} \equiv -\delta_t - x_{it}'\beta - \bar{x}_i'\phi' f_t - e_{it}$ is a subset of $[L_t, K_t]$.

For assumption S1, we permit heteroskedasitity of v_{it} at t dimension. In assumption S2, other than standard assumptions, we require $\overline{x}_i$ has no effect on the distribution of v_{it} once conditioned on x_{it}. Assumption S3 is the large support assumption for the special regressor.

Remark 3.2.3 The existence of special regressor depends on the context of empirical analysis, and it may not be easy to find such a regressor in some cases. For more discussions about the special regressor, see Honoré and Lewbel (2002) and Lewbel et al (2012).

Based on the above assumption, we have the following identification proposition, which is similar to Theorem 1 of Honore and Lewbel (2002).

Lemma 3.2.4 *Under assumptions S1, S2, and S3, let*

$$w_{it} = \frac{[y_{it} - 1\,(v_{it} > 0)]}{f_t(v_{it}\,|x_{it}\,)} \tag{3.2.8}$$

then we have

$$E(w_{it}\,|x_{i1}, \overline{x}_i\,) = \delta_t + x_{it}'\beta + \bar{x}_i'\psi' f_t \tag{3.2.9}$$

As a result, by introducing w_{it} and the special regressor v_{it}, we successfully transformed the nonlinear binary choice model into a linear model, and we will mainly consider the estimation of β based on the equation $E(w_{it}\,|x_{it}, \bar{x}_i\,) = \delta_t + x_{it}'\beta + \bar{x}_i'\psi' f_t$.

3.2.2 Identification

For binary choice model, it's always necessary to point out the identification condition for parameters of interest. Without further restriction, if the support of the observed predictor variables is bounded, then the binary choice model can only be identified in the logistic case (Chamberlain (2010)). However, the identification of special regressor approach is somewhat different from the usual approach. For special regressor approach, upon transformation, the nonlinear model has a linear representation (here, the linear representation is (3.2.9)). As a result, the identification of β can be achieved by applying the identification results of linear

partition regression. Let $\pi_i \equiv (1\ \overline{x}_i')'$, $g_t \equiv (\delta_t\ f_t'\psi)'$. Then the conditional expectation (3.2.9) for w_{it} could be rewritten as

$$E(w_{it}|x_{it},\overline{x}_i) = x_{it}'\beta + \pi_i' g_t = \left[x_{it}', \pi_i'\right]\begin{pmatrix}\beta \\ g_t\end{pmatrix} \equiv H_{it}\Lambda_t \tag{3.2.10}$$

where $H_{it} \equiv [x_{it}', \pi_i']$ and $\Lambda_t \equiv \left[\beta', g_t'\right]'$. As a result, the identification follows if the matrix $E\left[\boldsymbol{H}_t'\boldsymbol{H}_t\right]$ defined below is of full rank.

Assumption: (Identification) Let $\boldsymbol{H}_t = [H_{1t}', H_{2t}', \ldots, H_{Nt}']$, and $E\left[\boldsymbol{H}_t'\boldsymbol{H}_t\right]$ is of full rank.

It obvious that the assumption that $E\left[\boldsymbol{H}_t'\boldsymbol{H}_t\right]$ is of full rank requires $N \geq 2k+1$ because we have $2k+1$ unknown parameters in the model, we will maintain this implicit assumption throughout our paper.

Lemma 3.2.5 *Under assumptions 1, 2, S1, S2, S3 together with the above identification assumption, β is identified.*

The proof of this lemma is straightforward: by equation (3.2.10) from Lemma 3.2.4, Λ_t is identified by the full rank of $E\left[\boldsymbol{H}_t'\boldsymbol{H}_t\right]$, consequently β is identified.

Remark 3.2.6 The identification results of lemma (3.2.5) is similar to that of Lewbel (2000a), and is straightforward in that, once we transformed the binary choice model into the linear model, the identification results of linear model can be directly applied here.

3.2.3 Estimation

In the above section, we discuss how to transform the nonlinear binary choice panel data model into a possible linear regression model. In order to estimate the parameters of interest, β, we can apply the two-step estimation method. In the first step, we apply the

nonparametric method to estimate w_{it} of (3.2.8). In the second step, we apply OLS or GMM method to estimate β of (3.2.9). In the second step, to focus on the parameters of interest, we difference out the nuisance parameter g_t first. It would also be very desirable to do so, if we have some large T. This could be done in the following standard way. Let $A_t \equiv E(\pi_i \pi_i')^{-1} E(\pi_i x_{it}')$, $x_{it}^r \equiv x_{it} - A_t' \pi_i$, then we have

$$\begin{aligned} E\left(x_{it}^r w_{it}\right) &= E\left[x_{it}^r E\left(w_{it} | x_{it}, \overline{x}_i\right)\right] = E\left[x_{it}^r \left(x_{it}'\beta + \pi_i' g_t\right)\right] \\ &= E\left[\left(x_{it} - A_t'\pi_i\right)\left(x_{it}'\beta + \pi_i' g_t\right)\right] = E\left[x_{it}^r x_{it}^{r\prime}\right]\beta, \end{aligned}$$

and

$$\beta = E\left[x_{it}^r x_{it}^{r\prime}\right]^{-1} E\left(x_{it}^r w_{it}\right).$$

Therefore, our sample counterpart estimator could be

$$\begin{aligned} \widehat{\beta} &= \left(\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N} \widehat{x}_{it}^r \widehat{x}_{it}^{r\prime}\right)^{-1} \left(\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N} \widehat{x}_{it}^r \widehat{w}_{it}\right) \\ &= \left(\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N} \left(x_{it} - \widehat{A}_t'\pi_i\right)\left(x_{it} - \widehat{A}_t'\pi_i\right)'\right)^{-1} \left(\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N} \left(x_{it} - \widehat{A}_t'\pi_i\right)\widehat{w}_{it}\right), \end{aligned} \tag{3.2.11}$$

where $\widehat{w}_{it} = \frac{y_{it} - 1(v_{it} > 0)}{\widehat{f}(v_{it}|x_{it})}$ is a nonparametric estimate for w_{it}, and

$$\widehat{A}_t = \left(\frac{1}{N}\sum_{i=1}^{N} \pi_i \pi_i'\right)^{-1} \left(\frac{1}{N}\sum_{i=1}^{N} \pi_i x_{it}'\right).$$

A possible estimator of $f_t(v_{it}|x_{it})$ is standard Nadaraya-Watson estimator:

$$\widehat{f}_t(v_{it} | \mathbf{x}_{it}) = \frac{\widehat{f}_t(v_{it}, x_{it})}{\widehat{f}_t(x_{it})} = \frac{(N\boldsymbol{h})^{-1}\sum_{k=1}^{N} K_h\left(v_{kt} - v_{it}, x_{kt} - x_{it}\right)}{\left(N\widetilde{\boldsymbol{h}}\right)^{-1}\sum_{k=1}^{N} K_{\tilde{h}}\left(x_{kt} - x_{it}\right)} \tag{3.2.12}$$

where $\boldsymbol{h} = h_1 h_2 \cdots h_{k+1}$, $\widetilde{\boldsymbol{h}} = h_2 \cdots h_{k+1}$, $K_h(u) = \prod_{l=1}^{k+1} k\left(\frac{u_l}{h_l}\right)$, $K_{\tilde{h}}(u) = \prod_{l=2}^{k+1} k\left(\frac{u_l}{\tilde{h}_l}\right)$, $h = (h_1, \ldots, h_{k+1})'$, and $\tilde{h} = (h_2, \ldots, h_{k+1})'$. For simplicity, let $h_1 = h_2 = ... = h_{k+1} = h$. This simplification is just for theoretical convenience. In practice, one could use Silverman's rule of thumb to choose h_i, or use cross-validation method.

Given the above argument, we can estimate w_{it} by

$$\widehat{w}_{it} = \frac{[y_{it} - 1(v_{it} > 0)]}{\widehat{f}_t(v_{it} \,|x_{it}\,)} \tag{3.2.13}$$

where $\widehat{f}_t(v_{it} \,|x_{it}\,)$ is given by (3.2.12).

Remark 3.2.7 Recently, Dong and Lewbel (2012) propose a simple way to estimate w_{it} of equation (3.2.6), which starts with imposing assumptions on the special regressor, v_{it}, $V = S'b + U$, $E(U) = 0, U \perp S, \varepsilon$, $U \sim f(U)$, and then define T by $T = \frac{y_{it} - 1(v_{it} \geq 0)}{f(U)}$, which is equivalent to w_{it} in our framework. We focus on the method of Lewbel (2000a) since it's a more general approach.

3.2.4 Asymptotic analysis

Given the estimation of $\widehat{\beta}$, we are interested in looking at its limiting behavoir when $N \to \infty$. The asymptotic normality of $\widehat{\beta}$ is standard as in Newey and McFadden (1994). Though we have an estimated $\widehat{A}_t$, due to the root-N convergence of $\widehat{A}_t$ and nonparametric smoothing, the preliminary estimation of A_t has no impact on the final asymptotics.

Denote $\phi_{it} = (x_{it} - A_t'\pi_i)[y_{it} - 1(v_{it} > 0)]$, and let

$$\chi_{it} = (x_{it} - A_t'\pi_i)w_{it} = \frac{(x_{it} - A_t'\pi_i)[y_{it} - 1(v_{it} > 0)]}{f(v_{it}|x_{it})} = \frac{\phi_{it} f(x_{it})}{f(v_{it}, x_{it})}$$

Let

$$q_{it} = \chi_{it} - E(\chi_{it}|x_{it}, \pi_i) + E(\chi_{it}|x_{it}) - E(\chi_{it}|v_{it}, x_{it}). \tag{3.2.14}$$

To understand q_{it}, it is very similar to the q_{it} in Lewbel (2000a). Note that

$$E(\chi_{it}|x_{it}, \pi_i) = (x_{it} - A_t'\pi_i)E(w_{it}|x_{it}, \pi_i) = (x_{it} - A_t'\pi_i)(x_{it}'\beta + \pi_i'\gamma_t),$$

which plays the same role as $zx^T\beta$ in equation (4.12) in Lewbel (2000a).

Our main result is given in the following theorem, and its proof is provided in the appendix.

Theorem 3.2.8 *Under Assumption 1, S1, S2, S3, and technical assumptions in the appendix, let*

$$\Delta = \frac{1}{T}\sum_{t=1}^{T}[E(x_{it}x_{it}') - E(x_{it}\pi_i')E(\pi_i\pi_i')^{-1}E(\pi_i x_{it}')]$$

then the following holds,

$$\sqrt{N}(\widehat{\beta} - \beta) \xrightarrow{d} N\left(0, \Delta Var\left(\frac{1}{T}\sum_{t=1}^{T} q_{it}\right)\Delta'\right) \quad as\ N \to \infty \tag{3.2.15}$$

Assuming fixed T allows us to come out a clean asymptotics as in equation (3.2.15). From the proof of our lemmas and theorems, not hard to see that our results could be extended to the case when T goes to infinity without much difficulty, however the convergence rate might be different if T is larger than N. Assumption 2 and S2 are more likely to hold as T goes to infinity: $\overline{x}_i$ tends to the true underlying individual characteristics and $\overline{x}_i$ is less likely to have time effect, therefore even with strong serial correlation, once conditional on x_{it}, $\overline{x}_i$ is more likely not to affect the distribution of v_{it}.

Remark 3.2.9 For an consistent estimator of the variance term of the limiting distribution, we can replace Δ and $Var\left(\frac{1}{T}\sum_{t=1}^{T} q_{it}\right)$ by their sample counterpart estimators respectively. For example, we can replace Δ by $\widehat{\Delta} = \frac{1}{T}\sum_{t=1}^{T}[\widehat{E}(x_{it}x_{it}') - \widehat{E}(x_{it}\pi_i')\widehat{E}(\pi_i\pi_i')^{-1}\widehat{E}(\pi_i x_{it}')]$ where $\widehat{E}(A)$ denotes the estimator of $E(A)$ and usually the sample average. For the estimators of $Var\left(\frac{1}{T}\sum_{t=1}^{T} q_{it}\right)$, we could estimate it by

$$\widehat{Var}\left(\frac{1}{T}\sum_{t=1}^{T} q_{it}\right) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T}\widehat{q}_{it}\right)^2 - \left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\widehat{q}_{it}\right)^2$$

and $\widehat{q}_{it}$ is the estimator of q_{it} and it can be obtained by replacing the terms of (3.2.14) by corresponding nonparametric estimators (for example, kernel estimators) as follows

$$\widehat{q}_{it} = \widehat{\chi}_{it} - \widehat{E}(\chi_{it}|x_{it},\pi_i) + \widehat{E}(\chi_{it}|x_{it}) - \widehat{E}(\chi_{it}|v_{it},x_{it}).$$

3.2.5 Choice of special regressors

In this paper, special regressor methods assume that the model includes a single regressor, call it V, which has the following two properties. First the special regressor V is exogenous and additive to the model error, and then, the special regressor V is continuously distributed, and has a large support, so it can take on a wide range of values.[2,3] Details of special regressor methods can be found in Dong and Lewbel (2012), Lewbel (2012) and references therein.

[2] For example, any normally distributed regressor would automatically satisfy this continuous with large support condition.

[3] No matter how many endogenous regressors are in the model, only one special regressor that satisfies these properties is needed.

The remaining job is how to choose a special regressor. According to Lewbel et al (2012), other things equal, if there are more than one regressor in the model satisfies the required conditions to be special, in general the one with the thickest tails (e.g., having the largest variance) will typically be the best choice of special regressor, because it's desirable for efficiency and can affect rates of convergence.

3.3 Monte Carlo Simulation

In the above sections, we have established the asymptotic properties of the special regressor estimation of β. In this section, we conduct several experiments to check the performance of our proposed estimators. The design is as follows, and it is very close to the setting of Bai (2009b) in the linear panel data framework.

Model 1: Our first model has the form of $(r = 1)$

$$\begin{aligned} y_{it}^{*} &= \delta_t + v_{it} + \beta_1 x_{it} + \lambda_i f_t + \varepsilon_{it} \\ x_{it} &= 1 + \lambda_i f_t + \xi_{it} \\ y_{it} &= 1(y_{it}^{*} > 0) \end{aligned}$$

where $t = 1, ..., T; i = 1, ..., N$, with T is set to vary from 3, 5, 10, and N is set to vary from 50, 100, 500, 1000; $\beta_1 = 1$, $\delta_t = 0.9 - 0.2(t-1)$, λ_i, ξ_{it} are all iid $N(0,1)$, $f_t \sim_{iid} N(0,2), \varepsilon_{it} \sim_{iid} N(0, \sigma_i^2)$ with $\sigma_i^2 \sim \chi^2(1)$, $v_{it} \sim N(0,2)$, all of them are i.i.d. across i, and t. The simulation results are provided in Table 1.

Model 2: Our second model has the form of $(r = 1)$

$$\begin{aligned} y_{it}^{*} &= \delta_t + v_{it} + \beta_1 x_{it} + \lambda_i f_t + \varepsilon_{it} \\ x_{it} &= 1 + \lambda_i f_t + \xi_{it} \\ y_{it} &= 1(y_{it}^{*} > 0) \end{aligned}$$

where $t = 1, ..., T; i = 1, ..., N$, with T is set to vary from 3, 5, 10, and N is set to vary from 50, 100, 500, 1000; $\beta_1 = 1$, $\delta_t = 0.9 - 0.2(t-1)$, λ_i, ξ_{it} are all iid $N(0,1)$, $f_t \sim_{iid} N(0,2)$, $\varepsilon_{it} = \rho_i \varepsilon_{i,t-1} + \epsilon_{it}$ with $\epsilon_{it} \sim_{iid} N(0,2)$ and $\rho_i \sim IIDU\left[0.1, 0.9\right]$, $v_{it} \sim N(0,2)$, all of them are i.i.d. across i, and t. The simulation results are provided in Table 2.

Model 3: Our third model has the form of $(r = 1)$

$$\begin{aligned} y_{it}^{*} &= \delta_t + v_{it} + \beta_1 x_{it} + \lambda_i f_t + \varepsilon_{it} \\ x_{it} &= 1 + \lambda_i f_t + \xi_{it} \\ y_{it} &= 1(y_{it}^{*} > 0) \end{aligned}$$

where $t = 1, ..., T; i = 1, ..., N$, with T is set to vary from 3, 5, 10, and N is set to vary from 50, 100, 500, 1000; $\beta_1 = 1$, $\delta_t = 0.9 - 0.2(t-1), \lambda_i, \xi_{it}$ are all $iidN(0,1), f_t \sim_{iid} N(0,2)$, $\varepsilon_{it} = \sqrt{\chi_{it}}\epsilon_{it}$ with $\chi_{it} = 0.5 + x_{it}^2/20$ and $\epsilon_{it} \sim_{iid} N(0,1)$, $v_{it} \sim N(0,2)$, all of them are i.i.d. across i, and t. The simulation results are provided in Table 3.

For the above 3 DGPs, we assume the presence of one single interactive effects. The first one is the usual unconditional heterogenous variance model, the second one accommodates the case when the error is generated by stationary AR(1) process, and the last DGP considers conditional heterogeneous variance. These three cases are general enough

to accommodate the variability of economic situations.

For comparison in simulations, we compute the estimators of β using MLE_naive, MLE_infeasible and the special regressor method (trimmed and untrimmed). For the MLE_naive, it doesn't consider the interactive effects, i.e., only the regressors x_{it} are used in estimating β, and it is called naive estimator simply because it ignores the unobservable interactive effects in the model. The MLE_infeasible takes interactive effects into account, and treat the unobservable interactive effects as additional regressors. Hence it's the benchmark for comparison since MLE usually is the efficient one for a full model, and it's infeasible due to the presumed knowledge of unobservable interactive effects in the model.

From the simulation results, we can find that the MLE_naive is very unsatisfactory and there is huge bias for MLE_naive in the simulation. However, our method reaches our expectation in different settings including seral correlation and heterogeneity (both conditional and unconditional), and outperforms the MLE in the presence of interactive effects. The most important finding is that with the increase of N and T, the estimators using our method are very close to the efficient estimation method of MLE_infeasible. However, in practice, the prior knowledge of normally distributed error terms might be an obstacle to apply directly the MLE_infeasible, as a result, the method proposed in this paper would be preferrable in estimation with the presence of interactive effects.

Table 1: Simulation results for DGP1

N	T	3				5				10			
		β_{inf}	β_{naive}	β_{SR}	β_{trim}	β_{inf}	β_{naiv}	β_{SR}	β_{trim}	β_{inf}	β_{naive}	β_{SR}	β_{trim}
50	Estim	1.0524	1.3038	0.7491	0.7525	1.0570	1.1714	0.7031	0.7255	1.1786	1.4002	0.8354	0.8356
	Bias	0.0524	0.3038	-0.2509	-0.2475	0.0570	0.1714	-0.2969	-0.2745	0.1786	0.4002	-0.1646	-0.1644
	ABias	0.1900	0.3129	0.2712	0.2698	0.1386	0.1899	0.2996	0.2785	0.2042	0.4002	0.1692	0.1686
	RMSE	0.2463	0.3959	0.3212	0.3207	0.1817	0.2365	0.3267	0.3071	0.2553	0.4241	0.1931	0.1926
100	Estim	1.0249	1.1100	0.8201	0.8099	1.0003	1.5470	1.1251	1.1055	1.0844	1.5106	0.8548	0.8515
	Bias	0.0249	0.1100	-0.1799	-0.1901	0.0003	0.5470	0.1251	0.1055	0.0844	0.5106	-0.1452	-0.1485
	ABias	0.1316	0.1546	0.2520	0.2483	0.1048	0.5479	0.1635	0.1488	0.1092	0.5106	0.1418	0.1485
	RMSE	0.1727	0.2041	0.3041	0.2994	0.1344	0.5657	0.2032	0.1868	0.1412	0.5211	0.1491	0.1560
500	Estim	1.0013	1.5280	0.8025	0.7456	0.9931	1.3874	0.9262	0.9041	1.0500	1.3684	1.0210	0.9436
	Bias	0.0013	0.0528	-0.1975	-0.2544	-0.0069	0.3874	-0.0738	-0.0959	0.0500	0.3684	0.0210	-0.0564
	ABias	0.0677	0.5280	0.2555	0.2612	0.0531	0.3874	0.1127	0.1115	0.0581	0.3684	0.0732	0.0646
	RMSE	0.0851	0.5353	0.2965	0.2937	0.0680	0.3921	0.1373	0.1326	0.0717	0.3708	0.0943	0.0772
1000	Estim	0.9986	1.1826	0.9914	0.9356	0.9916	1.3571	1.1001	1.0494	1.0452	1.3737	1.0509	1.0253
	Bias	-0.0014	0.1826	-0.0086	-0.0644	-0.0084	0.3571	0.1001	0.0494	0.0452	0.3737	0.0509	0.0253
	ABias	0.0298	0.1826	0.1356	0.0818	0.0241	0.3571	0.1054	0.0598	0.0523	0.3737	0.0571	0.0354
	RMSE	0.0377	0.1868	0.1862	0.0993	0.0299	0.3585	0.1283	0.0749	0.0673	0.3751	0.0707	0.0441

Note: The simulation is replicated for 1000 times. True $\beta = 1$. Abias = absolute bias. RMSE = root mean square errors.

Table 2: Simulation results for DGP2

N	T	3				5				10			
		β_{inf}	β_{naive}	β_{SR}	β_{trim}	β_{inf}	β_{naive}	β_{SR}	β_{trim}	β_{inf}	β_{naive}	β_{SR}	β_{trim}
50	Estim	1.0377	1.4171	0.6383	0.6392	1.1018	1.6984	0.7570	0.7486	1.0579	1.3525	0.7838	0.7818
	Bias	0.0377	0.4171	-0.3617	-0.3608	0.1018	0.6984	-0.2430	-0.2514	0.0579	0.3525	-0.2162	-0.2181
	ABias	0.2781	0.4274	0.3844	0.3845	0.2738	0.6968	0.2806	0.2841	0.1497	0.3543	0.2185	0.2204
	RMSE	0.3789	0.5571	0.4473	0.4488	0.3682	0.7768	0.3349	0.3389	0.1924	0.3974	0.2398	0.2411
100	Estim	1.0069	1.1393	0.8456	0.8283	1.0162	1.4061	0.8778	0.8858	1.0220	1.4335	1.0335	1.0333
	Bias	0.0069	0.1393	-0.1544	-0.1717	0.0162	0.4061	-0.1222	-0.1142	0.0220	0.4335	0.0335	0.0333
	ABias	0.1505	0.1895	0.2362	0.2352	0.1366	0.4062	0.1536	0.1499	0.0918	0.4335	0.0710	0.0713
	RMSE	0.1951	0.2431	0.2907	0.2904	0.1751	0.4403	0.1860	0.1826	0.1175	0.4499	0.0904	0.0905
500	Estim	1.0091	1.5844	0.7874	0.7522	0.9957	1.4053	0.9261	0.9011	1.0416	1.3890	1.0235	0.9981
	Bias	0.0091	0.5844	-0.2126	-0.2478	-0.0043	0.4053	-0.739	-0.0989	0.0416	0.3890	0.0235	-0.0019
	ABias	0.0803	0.5844	0.2413	0.2565	0.0607	0.4053	0.1169	0.1158	0.0577	0.3890	0.0618	0.0489
	RMSE	0.1012	0.5940	0.2840	0.2931	0.0758	0.4118	0.1408	0.1396	0.0727	0.3923	0.0787	0.0619
1000	Estim	1.0041	1.3072	1.0644	1.0233	1.0001	1.4943	0.9773	0.9539	1.0334	1.4175	1.0483	1.0354
	Bias	0.0041	0.3072	0.0644	0.0233	0.0001	0.4943	-0.0227	-0.0461	0.0334	0.4175	0.0483	0.0354
	ABias	0.0483	0.3072	0.1049	0.0735	0.0408	0.4943	0.0693	0.0653	0.0445	0.4175	0.0544	0.0436
	RMSE	0.0617	0.3129	0.1422	0.0937	0.0508	0.4971	0.0906	0.0809	0.0569	0.4197	0.0670	0.0547

Note: The simulation is replicated for 1000 times. True $\beta = 1$. Abias = absolute bias. RMSE = root mean square errors.

Table 3: Simulation results for DGP3

N	T	3				5				10			
		β_{inf}	β_{naive}	β_{SR}	β_{trim}	β_{inf}	β_{naive}	β_{SR}	β_{trim}	β_{inf}	β_{naive}	β_{SR}	β_{trim}
50	Estim	0.9926	1.3494	0.6754	0.6769	0.9838	1.2241	1.0257	1.0351	1.0891	1.3065	0.8225	0.8225
	Bias	-0.0074	0.3494	-0.3246	-0.3231	-0.0162	0.2241	0.0257	0.0351	0.0891	0.3065	-0.1775	-0.1775
	ABias	0.2318	0.3636	0.3458	0.3450	0.1784	0.2416	0.1295	0.1315	0.1513	0.3070	0.1810	0.1798
	RMSE	0.2968	0.4511	0.4062	0.4071	0.2229	0.3031	0.1624	0.1641	0.1988	0.3363	0.2014	0.2001
100	Estim	0.9659	1.1021	0.8662	0.8495	0.9474	1.3592	0.9090	0.9163	1.0004	1.3900	1.0770	1.0780
	Bias	-0.0351	0.1021	-0.1338	-0.1505	-0.0526	0.3592	-0.0910	-0.0837	0.0004	0.3900	0.0770	0.0780
	ABias	0.1334	0.1548	0.2236	0.2230	0.1187	0.3592	0.1323	0.1301	0.0786	0.3900	0.0899	0.0909
	RMSE	0.1667	0.1981	0.2733	0.2730	0.1473	0.3836	0.1623	0.1598	0.0975	0.4010	0.1099	0.1109
500	Estim	0.9487	1.4830	0.7951	0.7704	0.9396	1.3038	0.9568	0.9381	1.0218	1.3079	1.0449	1.0254
	Bias	-0.0513	0.4830	-0.2049	-0.2296	-0.0604	0.3038	-0.0432	-0.0619	0.0218	0.3079	0.0449	0.0254
	ABias	0.0752	0.4830	0.2278	0.2373	0.0718	0.3038	0.0998	0.0937	0.0487	0.3079	0.0626	0.0485
	RMSE	0.0937	0.4895	0.2640	0.2751	0.0867	0.3082	0.1252	0.1132	0.0647	0.3102	0.0809	0.0716
1000	Estim	0.9450	1.2094	1.0790	1.0448	0.9404	1.4253	0.9899	0.9747	1.0004	1.3527	1.0727	1.0638
	Bias	-0.0550	0.2094	0.0790	0.0448	-0.0596	0.4253	-0.0101	-0.0253	0.0004	0.3527	0.0727	0.0638
	ABias	0.0613	0.2094	0.1041	0.0745	0.0619	0.4253	0.0628	0.0575	0.0371	0.3527	0.0732	0.0644
	RMSE	0.0734	0.2147	0.1407	0.0946	0.0716	0.4272	0.0802	0.0717	0.0472	0.3538	0.0739	0.0683

Note: The simulation is replicated for 1000 times. True $\beta = 1$. Abias = absolute bias. RMSE = root mean square errors.

3.4 Empirical application

In order to apply our special regressor method to empirical studies, we consider the women's laborforce participation. The data contains 1461 married women of the Panel Study of Income Dynamics (PSID) for 10 calendar years 1979–1988[4]. The women's laborforce participation has been widely studied by econometricians. First, Hyslop (1999) considers a dynamic search framework to analyze the intertemporal labor force participation behavior of married women, where he considers linear probability and probit models and the dynamic probit models are estimated using maximum simulated likelihood (SML) estimation. After that, Carro (2007) applies the similar data using the modified maximum likelihood in a dynamic setting. More recently, Wooldridge (2010) employs a panel data model for women's laborforce participation, where he assumes the error term is normally distributed.

To summarize, most of the researches on women's labor force participation assumes normally distributed error term and use large cross section data with small fixed time period. As pointed out in the introduction, our estimation approach adapts these situations very well because we don't require the errors to be normally distributed and we assume T is usually small. As a result, we will consider the following model for women's labor force participation,

$$y_{it} = 1\left(v_{it} + \delta_t + x_{it}'\beta + \lambda_i f_t + \varepsilon_{it}\right)$$

where y_{it} takes value one if women i participate in period t and zero otherwise, $x_{it} =$ (#children0-2$_{it}$, #children3-5$_{it}$, #children6-17$_{it}$, log$income_{it}$, time effect, race), where #children$a$-b is the number of children aged between a and b, log$income$ is the log of husband's labor income deflated by Consumer Price Index and age is wife's age. These variables are con-

[4]We appreciate Dr. Carro very much for generously providing us the data for analysis.

sidered by Carro (2007) as well as Hyslop (1999) and Wooldridge (2010). In the researches of Carro (2007) and Hyslop (1999), they allow time dummy variables to specify the time effects, which can be interpreted as time effect δ_t in our set up.

For the choice of special regressor, we use the negative age minus the whole sample mean as the special regressor, the transformation is to make sure that age has a positive coefficient and zero mean. This is suggested by Dong and Lewbel (2012) and by the fact that the current researches suggest that the estimated coefficient of age is significant negative (for example, Hyslop (1999) and Wooldridge (2010)).

For the women's labor force participation analysis, as pointed by Hyslop (1999), there is so called "*taste of work*" which is unobservable and affects the labor force participation. Moreover, this "*taste of work*" is correlated with the realization of fertility as well as non-labor income. As a result, to take into account of this effects, we can use λ_i to denote the "*taste of work*", and will use the Mundlak's projection method to approximate this taste, i.e.,

$$E\left(\lambda_i | x_{i1}, \cdots, x_{iT}\right) = \lambda + \phi \bar{x}_i$$

with $\bar{x}_i = \frac{1}{T}\sum_{t=1}^{T} x_{it}$ where x_{it} are the observable variables and are given above. As a result, f_t can interpreted as the time effects of "*taste of work*" at different time.

For comparison, we consider two probit models. One is the same as above, assuming that ε_{it} is standard normal. A simple alternative is as follows

$$\begin{aligned} y_{it}^* &= v_{it} + \delta_t + x_{it}'\beta + \sigma\varepsilon_{it}, \\ y_{it} &= 1(y_{it}^* \geqslant 0), \end{aligned}$$

we can note that for the complete model, we need to estimate $T + 4 + 3T + T = 54$ parameters, which is a lot. However, for the simple alternative model, we only need to estimate $T + 4 + 1 = 15$ parameters.

In order to apply our estimation approach for analyzing women's labor force participation, we use the normal kernel density for nonparametric estimation, and choose the bandwidth by Silverman's Rule-of-Thumb. Of course, the optimal choice of kernel density and associated bandwidth is beyond the scope of current paper. Below, only the estimates of β is reported.

Table 4: Estimation results for women's labor force participation

	Special Regressor	Probit (complete)	Probit (simple)
child0_2	-0.855	-3.755	-2.014
	(0.187)	(0.034)	(0.088)
child3_5	-0.601	-2.061	-1.318
	(0.198)	(0.039)	(0.097)
child6_17	0.205	0.091	-0.098
	(0.529)	(0.099)	(0.234)
logincome	-0.113	0.039	0.304
	(0.039)	(0.687)	(1.644)
race	0.063	0.304	0.615
	(0.110)	(1.644)	(0.069)

From the above table, several interesting results can be found. The main finding is that we find that husbands' income has positive significant effect on women's labor force participation. This result is consistent with the finding of Carro (2007) in dynamic setting, and suggests that husbands' income should have negative significant effect on women's labor force participation. However, there would be no significant effects of on women's

labor force participation if we apply the probit model, which is adapted by Wooldridge (2012). As mentioned above, this results is counter-intuitive since it's normal that married women are not willing to participate work if the husbands' income is high. It's obvious that the special regressor method proposed in this paper capture this and none of the other methods could obtain the similar results. All of these suggest that it would be inappropriate to apply the probit model when the data contains potentially unobserved interactive effects especially with a short time span, and that it would be appropriate to use our proposed estimation for taking into account of the unobserved interactive effects without presuming specific distributional assumptions.

3.5 Conclusion

In this paper, we consider the estimation of binary choice model with interactive effects through the special regressor approach. Since the interactive effects are usually unobservable, it would be problematic in modelling if they are ignored. To control the unobserved interactive effects, we adopt the Mundluk (1978)'s projection method, which uses projection method to model the unobserved interactive effects. Furthermore, we apply the special regressor method of Lewbel (2000), which transform the binary choice model into a linear model with the help of the so called special regressor.

Mento Carlo simulations show us that the special regressor estimator in our paper outperforms the MLE if the unobserved interactive effects are ignored in probit model, and this suggests that the special regressor estimator would be appropriate for modelling binary choice model with interactive effects and thus eliminating source of bias in binary choice model. Finally, we apply our model to analyze women's labor force participation. Com-

pared to the existing researches on women's labor force participation, the special regressor estimation results suggest that husbands' income should have negative significant effect on women's labor force participation, which is intuitive and consistent with the real world. Our next step is to apply the special regressor method to dynamic binary choice model, but this is beyond the current scope.

3.6 Appendix

3.6.1 Proof of lemma (3.2.4)

Proof.2 Note that $s_{it} \equiv -\delta_t - x_{it}'\beta - \bar{x}_i'\phi' f_t - e_{it}$,

$$
\begin{aligned}
& E(w_{it}\,|x_{it},\overline{x}_i\,) = E\left(\frac{[y_{it}-1\,(v_{it}>0)]}{f_t(v_{it}\,|x_{it}\,)}\,|x_{it},\overline{x}_i\right) \\
= \;& E\left(\frac{E\,[y_{it}-1\,(v_{it}>0)\,|v_{it},x_{it},\overline{x}_i]}{f_t(v_{it}\,|x_{it}\,)}\,|x_{it},\overline{x}_i\right) \\
= \;& \int_{L_t}^{K_t}\frac{E\,[y_{it}-1\,(v_{it}>0)\,|v_{it},x_{it},\overline{x}_i]}{f_t(v_{it}\,|x_{it}\,)}f_t(v_{it}\,|x_{it}\,)dv_{it} \\
= \;& \int_{L_t}^{K_t}\int_{\Omega_{e_{it}}}[1\,(v_{it}-s_{it}>0)-1\,(v_{it}>0)]\,dF_{e_{it}}\,(e_{it}\,|v_{it},x_{it,},\overline{x}_i\,)\,dv_{it} \\
= \;& \int_{\Omega_{e_{it}}}\int_{L_t}^{K_t}[1\,(v_{it}>s_{it})-1\,(v_{it}>0)]\,dv_{it}dF_{e_{it}}\,(e_{it}\,|x_{it},\overline{x}_i\,) \\
= \;& \int_{\Omega_{e_{it}}}\int_{L_t}^{K_t}[(1\,(v_{it}>s_{it})-1\,(v_{it}>0))\,1\,(s_{it}\le 0) \\
& +(1\,(v_{it}>s_{it})-1\,(v_{it}>0))\,1\,(s_{it}>0)]\,dv_{it}dF_{e_{it}}\,(e_{it}\,|x_{it},\overline{x}_i\,) \\
= \;& \int_{\Omega_{e_{it}}}\int_{L_t}^{K_t}[1\,(s_{it}<v_{it}\le 0)\,1\,(s_{it}\le 0)+1\,(0<v_{it}\le s_{it})\,1\,(s_{it}>0)]\,dv_{it}dF_{e_{it}}\,(e_{it}\,|x_{it},\overline{x}_i\,) \\
= \;& \int_{\Omega_{e_{it}}}\left[1\,(s_{it}\le 0)\int_{s_{it}}^{0}1dv_{it}-1\,(s_{it}>0)\int_{0}^{s_{it}}1dv_{it}\right]dF_{e_{it}}\,(e_{it}\,|x_{it},\overline{x}_i\,) \\
= \;& \int_{\Omega_{e_{it}}}-s_{it}dF_{e_{it}}\,(e_{it}\,|x_{it},\overline{x}_i\,) = \int_{\Omega_{e_{it}}}\left(\delta_t+x_{it}'\beta+\bar{x}_i'\psi' f_t+e_{it}\right)dF_{e_{it}}\,(e_{it}\,|x_{it},\overline{x}_i\,) \\
= \;& \delta_t+x_{it}'\beta+\bar{x}_i'\psi' f_t+E\,(e_{it}\,|x_{it},\overline{x}_i\,) = \delta_t+x_{it}'\beta+\bar{x}_i'\psi' f_t,
\end{aligned}
$$

where third and fifth line holds by $f_t\,(v_{it}|x_{it},\overline{x}_i)$ and $F_{e_{it}}\,(e_{it}\,|v_{it},x_{it,},\overline{x}_i\,) = F_{e_{it}}\,(e_{it}\,|x_{it,},\overline{x}_i\,)$ respectively, and the last line holds by $E\,(e_{it}\,|x_{it},\overline{x}_i\,) = 0$. ■

3.6.2 Proof of the theorem (3.2.8)

In order to analyze the limiting behavior of $\widehat{\beta}$, we impose several additional assumptions in the following.

Assumption A.1: ϕ_{it}, $f(x_{it})$ and $f(v_{it}, x_{it})$ are bounded, and $f(v_{it}, x_{it})$ is bounded away from zero.

Assumption A.2: There exist some functions $m_1(x)$, $m_2(v, x)$, $m_3(x)$, and $m_4(v, x)$ such that density function $f(x_{it})$, $f(v_{it}, x_{it})$, $E(\chi_{it}|x_{it})$, and $E(\chi_{it}|v_{it}, x_{it})$ satisfy the following local Lipschitz condition:

$$|f(x_{it} + c_x) - f(x_{it})| \leq m_1(x_{it}) \|c_x\|,$$

$$|f(v_{it} + c_v, x_{it} + c_x) - f(v_{it}, x_{it})| \leq m_2(v_{it}, x_{it}) \|(c_v, c_x)\|,$$

$$|E(\chi_{it}|x_{it} + c_x) - E(\chi_{it}|x_{it})| \leq m_3(x_{it}) \|c_x\|,$$

$$|E(\chi_{it}|v_{it} + c_v, x_{it} + c_x) - E(\chi_{it}|x_{it})| \leq m_4(v_{it}, x_{it}) \|(c_v, c_x)\|.$$

Also $E[m_1(x_{it})^2]$, $E[m_2(v_{it}, x_{it})^2]$, $E[m_3(x_{it})^2]$, and $E[m_4(v_{it}, x_{it})^2]$ exist.

Assumption A.3: The kernel functions $K(v, x)$ and $K(x)$ have supports that are convex on $\mathbb{R}^{k+1}$ and $\mathbb{R}^k$ respectively. $\int K(x)^2 dx$, $\int K(v, x)^2 dv dx$, $\int \|x\| K(x) dx$, and $\int \|(v, x)\| K(v, x) dv dx$ are finite. Both kernel functions are symmetric about zero and have order of p, which is

$$\int x_1^{l_1} ... x_k^{l_k} K(x) dx = 0 \quad \text{for } l_1 + ... + l_k < p,$$

$$\int x_1^{l_1} ... x_k^{l_k} K(x) dx \neq 0 \quad \text{for some } l_1 + ... + l_k = p.$$

This similarly holds for $K(v, x)$.

Assumption A.4: $h \to 0$, as $N \to \infty$, and there exists a small $\varepsilon > 0$, such that $N^{1-\varepsilon} h^{2(k+1)} \to \infty$, $N h^{2p} \to 0$.

Lemma 3.6.1 *Under Assumption A.1 to A.4, the following hold:*

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{\phi_{it}(\widehat{f}(x_{it})-f(x_{it}))}{f(v_{it},x_{it})}=\frac{1}{\sqrt{N}}\sum_{i=1}^{N}[E(\chi_{it}|x_{it})-E(\chi_{it})]+o_p(1),$$

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{\phi_{it}f(x_{it})(\widehat{f}(v_{it},x_{it})-f(v_{it},x_{it}))}{f^2(v_{it},x_{it})}=\frac{1}{\sqrt{N}}\sum_{i=1}^{N}[E(\chi_{it}|v_{it},x_{it})-E(\chi_{it})]+o_p(1).$$

Proof.2 The proof is a simple version of theorem B in Lewbel (2000b).

$\widehat{f}(x_{it})$ here is a leave-one-out nonparametric estimate, which is

$$\widehat{f}(x_{it})=\frac{1}{N-1}\sum_{j=1,j\neq i}^{N}\frac{1}{h^k}K\left(\frac{x_{it}-x_{jt}}{h}\right).$$

Let $\widehat{\mu}=\frac{1}{N}\sum_{i=1}^{N}\frac{\phi_{it}\widehat{f}(x_{it})}{f(v_{it},x_{it})}$, then

$$\widehat{\mu}=\frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N}\frac{1}{h^k}\frac{\phi_{it}}{f(v_{it},x_{it})}K\left(\frac{x_{it}-x_{jt}}{h}\right).$$

Since $K(x)$ is symmetric,

$$\widehat{\mu}=\frac{2}{N(N-1)}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\frac{1}{2h^k}\left(\frac{\phi_{it}}{f(v_{it},x_{it})}+\frac{\phi_{jt}}{f(v_{jt},x_{jt})}\right)K\left(\frac{x_{it}-x_{jt}}{h}\right).$$

Define $P(z_{it},z_{jt})$ by

$$P(z_{it},z_{jt})=\frac{1}{2h^k}\left(\frac{\phi_{it}}{f(v_{it},x_{it})}+\frac{\phi_{jt}}{f(v_{jt},x_{jt})}\right)K\left(\frac{x_{it}-x_{jt}}{h}\right),$$

Where $z_{it}=[\frac{\phi_{it}}{f(v_{it},x_{it})},x_{it}]$. The aymptotic property of $\widehat{\mu}$ follows from Lemma 3.1 in Powell, Stock, and Stoker (1989). To apply the lemma, we first need to prove that $E\left[\|P(z_{it},z_{jt})\|^2\right]=$

$O(N)$.

$$
\begin{aligned}
E\left[\|P(z_{it},z_{jt})\|^2\right] &= \iint \frac{1}{4h^{2k}} E\left[\left(\frac{\phi_{it}}{f(v_{it},x_{it})}+\frac{\phi_{jt}}{f(v_{jt},x_{jt})}\right)^2\middle| x_{it},x_{jt}\right] \\
&\quad K\left(\frac{x_{it}-x_{jt}}{h}\right)^2 f(x_{it})f(x_{jt})dx_{it}dx_{jt} \\
&\leq M_1 \iint \frac{1}{4h^{2k}} K\left(\frac{x_{it}-x_{jt}}{h}\right)^2 f(x_{it})f(x_{jt})dx_{it}dx_{jt},
\end{aligned} \tag{3.6.1}
$$

where M_1 is a sufficiently large number which could bound $E\left[\left(\frac{\phi_{it}}{f(v_{it},x_{it})}+\frac{\phi_{jt}}{f(v_{jt},x_{jt})}\right)^2\middle| x_{it},x_{jt}\right]$. The existence of M_1 is guaranteed by Assumption A.1. By changing variable $u_{it}=\frac{x_{it}-x_{jt}}{h}$,

$$
\iint \frac{1}{4h^{2k}} K\left(\frac{x_{it}-x_{jt}}{h}\right)^2 f(x_{it})f(x_{jt})dx_{it}dx_{jt} = \iint \frac{1}{4h^{k}} K\left(u_{it}\right)^2 f(x_{jt}+u_{it}h)f(x_{jt})du_{it}dx_{jt}.
$$

Since $\int K(x)^2dx$ is finite, the term above is $O(\frac{1}{h^k})$. By Assumption A.4, we know the above term is $O(N)$. Therefore, $E\left[\|P_{ij}\|^2\right]$ is $O(N)$. The preliminary conditions of Lemma 3.1 in Powell, Stock, and Stoker (1989) is thus satisfied, so the following holds:

$$
N^{\frac{1}{2}}[\widehat{\mu}-E(\widehat{\mu})] = N^{\frac{1}{2}}\sum_{i=1}^{N} 2\left[E(p(z_{it},z_{jt})|z_{it})-E(p(z_{it},z_{jt}))\right]+o_p(1).
$$

The term $2E(p(z_{it},z_{jt})|z_{it})$ is not clear at first glance. It could be written as follows:

$$
\begin{aligned}
2E(p(z_{it},z_{jt})|z_{it}) &= \int \frac{1}{h^k}\left[\frac{\phi_{it}}{f(v_{it},x_{it})}+E\left(\frac{\phi_{jt}}{f(v_{jt},x_{jt})}\middle| x_{jt}\right)\right] K\left(\frac{x_{it}-x_{jt}}{h}\right) f(x_{jt})dx_{jt} \\
&= \int \left[\frac{\phi_{it}}{f(v_{it},x_{it})}+E\left(\frac{\phi_{it}}{f(v_{it},x_{it}+hu)}\middle| x_{it}+hu\right)\right] K\left(u\right) f(x_{it}+hu)du. \\
&= \chi_{it}+E(\chi_{it}|x_{it})+\int \frac{\phi_{it}}{f(v_{it},x_{it})}[f(x_{it}+hu)-f(x_{it})]K\left(u\right)du \\
&\quad +\int\left[E\left(\frac{\phi_{it}f(x_{it}+hu)}{f(v_{it},x_{it}+hu)}\middle| x_{it}+hu\right)-E(\chi_{it}|x_{it})\right]K(u)du.
\end{aligned}
$$

Let

$$\begin{aligned}\varsigma_{it} &= 2E[p(z_{it}, z_{jt})|z_{it}) - \chi_{it} - E(\chi_{it}|x_{it})] \\ &= \int \left\{ \frac{\phi_{it}}{f(v_{it}, x_{it})}[f(x_{it}+hu) - f(x_{it})] - \left[E\left(\frac{\phi_{it} f(x_{it}+hu)}{f(v_{it}, x_{it}+hu)} \middle| x_{it}+hu \right) - E(\chi_{it}|x_{it}) \right] \right\} K(u)\, du,\end{aligned}$$

then

$$\begin{aligned}N^{\frac{1}{2}}[\widehat{\mu} - E(\widehat{\mu})] &= N^{\frac{1}{2}} \sum_{i=1}^{N} \{\chi_{it} + E(\chi_{it}|x_{it}) - E[\chi_{it} + E(\chi_{it}|x_{it})]\} \\ &\quad + N^{\frac{1}{2}} \sum_{i=1}^{N} (\varsigma_{it} - E(\varsigma_{it})) + o_p(1).\end{aligned} \tag{3.6.2}$$

Using local Lipschitz conditions in Assumption A.2, ς_{it} is $O_p(h)$, and

$$E(\varsigma_{it}^2) \leq h^2 E\left[\left(\frac{\phi_{it}}{f(v_{it}, x_{it})} m_1(x_{it}) + m_3(x_{it}) \right)^2 \right] \left[\int \|u\| K(u) du \right]^2 = O(h^2) = o(1),$$

which implies that $N^{\frac{1}{2}} \sum_{i=1}^{N} (\varsigma_{it} - E(\varsigma_{it}))$ is $o_p(1)$, by Assumption A.1 and A.2.

For $E(\widehat{\mu})$,

$$\begin{aligned}E(\widehat{\mu}) &= E\left[\frac{1}{h^k} \frac{\phi_{it}}{f(v_{it}, x_{it})} K\left(\frac{x_{it} - x_{jt}}{h} \right) \right] \\ &= \iint \frac{1}{h^k} E\left(\frac{\phi_{it}}{f(v_{it}, x_{it})} \middle| x_{it} \right) K\left(\frac{x_{it} - x_{jt}}{h} \right) f(x_{it}) f(x_{jt}) dx_{it} dx_{jt} \\ &= \iint \frac{1}{h^k} E\left(\chi_{it} | \, x_{it} \right) K\left(\frac{x_{it} - x_{jt}}{h} \right) f(x_{jt}) dx_{it} dx_{jt} \\ &= \iint E\left(\chi_{it} | \, x_{it} \right) K\left(u_{jt} \right) f(x_{it} + h u_{jt}) dx_{it} du_{jt}.\end{aligned}$$

Since $K(x)$ is p-th order kernel,

$$\begin{aligned} E(\widehat{\mu}) &= \int E(\chi_{it}|\, x_{it})\, f(x_{it}) dx_{it} + O_p(h^p) \\ &= E(\chi_{it}) + O_p(h^p). \end{aligned} \tag{3.6.3}$$

By Assumption A.4, equation (3.6.2), and equation (3.6.3),

$$N^{\frac{1}{2}}[\widehat{\mu} - E(\chi_{it})] = N^{\frac{1}{2}} \sum_{i=1}^{N} \{\chi_{it} + E(\chi_{it}|x_{it}) - E[\chi_{it} + E(\chi_{it}|x_{it})]\} + o_p(1).$$

Reorganize it, we have

$$N^{\frac{1}{2}}\widehat{\mu} = N^{\frac{1}{2}} \sum_{i=1}^{N} [\chi_{it} + E(\chi_{it}|x_{it}) - E(\chi_{it})] + o_p(1).$$

Move χ_{it} from right-hand side to left-hand side, then it is the first conclusion in this lemma.

The second conclusion follows similarly. ■

Proof of Theorem (3.2.8).2 Since

$$\frac{1}{NT} \sum_{t=1}^{T} \sum_{i=1}^{N} \left(x_{it} - \widehat{A}_t' \pi_i \right) \pi_i' = 0,$$

and

$$\frac{1}{NT} \sum_{t=1}^{T} \sum_{i=1}^{N} \left(x_{it} - \widehat{A}_t' \pi_i \right) \left(x_{it} - \widehat{A}_t' \pi_i \right)' = \frac{1}{NT} \sum_{t=1}^{T} \sum_{i=1}^{N} \left(x_{it} - \widehat{A}_t' \pi_i \right) x_{it}',$$

then the following hold

$$\widehat{\beta} - \beta = \left[\frac{1}{NT} \sum_{t=1}^{T} \sum_{i=1}^{N} \left(x_{it} - \widehat{A}_t' \pi_i \right) \left(x_{it} - \widehat{A}_t' \pi_i \right)' \right]^{-1} \left[\frac{1}{NT} \sum_{t=1}^{T} \sum_{i=1}^{N} \left(x_{it} - \widehat{A}_t' \pi_i \right) \left(\widehat{w}_{it} - x_{it}' \beta - \pi_i' \lambda_t \right) \right].$$

Remeber that $E(w_{it}|x_{it},\pi_i) = x'_{it}\beta + \pi'_i\lambda_t$, so

$$\widehat{\beta}-\beta = \left[\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\left(x_{it}-\widehat{A}'_t\pi_i\right)\left(x_{it}-\widehat{A}'_t\pi_i\right)'\right]^{-1}\left[\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\left(x_{it}-\widehat{A}'_t\pi_i\right)\left(\widehat{w}_{it}-E(w_{it}|x_{it},\pi_i)\right)\right].$$

For the first term in $\widehat{\beta}-\beta$,

$$\begin{aligned}&\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\left(x_{it}-\widehat{A}'_t x_{it}\right)\left(x_{it}-\widehat{A}'_t x_{it}\right)' \qquad (3.6.4)\\ =\ &\frac{1}{T}\sum_{t=1}^{T}\left[\frac{1}{N}\sum_{i=1}^{N}x_{it}x'_{it}-\left(\frac{1}{N}\sum_{i=1}^{N}x_{it}\pi'_i\right)\left(\frac{1}{N}\sum_{i=1}^{N}\pi_i\pi'_i\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\pi'_i x_{it}\right)\right],\end{aligned}$$

We do not impose any assumptions about relationship across t. Under the condition that observations are i.i.d. across dimension i ,

$$\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\left(x_{it}-\widehat{A}'_t\pi_i\right)\left(x_{it}-\widehat{A}'_t\pi_i\right)' \xrightarrow{p} \frac{1}{T}\sum_{t=1}^{T}\left[E(x_{it}x'_{it})-E(x_{it}\pi'_i)E(\pi_i\pi'_i)^{-1}E(\pi'_i x_{it})\right], \qquad (3.6.5)$$

and remember the notation made in the theorem $\Delta = \frac{1}{T}\sum_{t=1}^{T}[E(x_{it}x'_{it})-E(x_{it}\pi'_i)E(\pi_i\pi'_i)^{-1}E(\pi'_i x_{it})]$, so

$$\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}(x_{it}-\widehat{A}'_t\pi_i)(x_{it}-\widehat{A}'_t\pi_i)' \xrightarrow{p} \Delta. \qquad (3.6.6)$$

For the second term in $\widehat{\beta}-\beta$, multiply it by $\sqrt{N}$,

$$\begin{aligned}&\sqrt{N}\left[\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\left(x_{it}-\widehat{A}'_t\pi_i\right)\left(\widehat{w}_{it}-E(w_{it}|x_{it},\pi_i)\right)\right]\\ =\ &\frac{1}{\sqrt{N}T}\sum_{t=1}^{T}\sum_{i=1}^{N}\left(x_{it}-A'_t\pi_i\right)\left(\widehat{w}_{it}-E(w_{it}|x_{it},\pi_i)\right)-\frac{1}{\sqrt{N}T}\sum_{t=1}^{T}\left(\widehat{A}_t-A_t\right)'\sum_{i=1}^{N}\pi_i(\widehat{w}_{it}-E(w_{it}|x_{it},\pi_i)).\end{aligned}$$

$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\pi_i(\widehat{w}_{it}-E(w_{it}|x_{it},\pi_i))$ could be shown to be $O_p(1)$ and since $(\widehat{A}_t-A_t)$ is $o_p(1)$ and

T is fixed, the following holds

$$\frac{1}{\sqrt{N}T}\sum_{t=1}^{T}\left(\widehat{A}_t - A_t\right)'\sum_{i=1}^{N}\pi_i(\widehat{w}_{it} - E(w_{it}|x_{it},\pi_i)) = o_p(1).$$

Therefore, the influence function for the second term in $\widehat{\beta} - \beta$ is $\frac{1}{\sqrt{N}T}\sum_{t=1}^{T}\sum_{i=1}^{N}(x_{it} - A_t'\pi_i)(\widehat{w}_{it} - E(w_{it}|x_{it},\pi_i))$. Since A_t is a constant, the influence function becomes a special case of Lewbel (2000a) and Honoré and Lewbel (2002).

If we know w_{it} and do not need to estimate them, then the asymptotic property of this part is straightforward. Remember the notation $\phi_{it} = (x_{it} - A_t'\pi_i)[y_{it} - 1(v_{it} > 0)]$ and $\chi_{it} = \frac{\phi_{it}f(x_{it})}{f(v_{it},x_{it})}$ then

$$\begin{aligned}
& \frac{1}{\sqrt{N}T}\sum_{t=1}^{T}\sum_{i=1}^{N}(x_{it} - A_t'\pi_i)(\widehat{w}_{it} - E(w_{it}|x_{it},\pi_i)) \\
= \; & \frac{1}{\sqrt{N}T}\sum_{t=1}^{T}\sum_{i=1}^{N}\left[\frac{\phi_{it}f(x_{it})}{f(v_{it},x_{it})} - (x_{it} - A_t'\pi_i)E(w_{it}|x_{it},\pi_i) + \frac{\phi_{it}(\widehat{f}(x_{it}) - f(x_{it}))}{f(v_{it},x_{it})}\right. \\
& \left. - \frac{\phi_{it}f(x_{it})(\widehat{f}(v_{it},x_{it}) - f(v_{it},x_{it}))}{f^2(v_{it},x_{it})} + R_{it}\right]
\end{aligned}$$

where

$$\begin{aligned}
R_{it} \; \equiv \; & \frac{\phi_{it}f(x_{it})(\widehat{f}(v_{it},x_{it}) - f(v_{it},x_{it}))}{f^2(v_{it},x_{it})} - \frac{\phi_{it}\widehat{f}(x_{it})(\widehat{f}(v_{it},x_{it}) - f(v_{it},x_{it}))}{f(v_{it},x_{it})\widehat{f}(v_{it},x_{it})} \\
= \; & \frac{\phi_{it}f(x_{it})(\widehat{f}(v_{it},x_{it}) - f(v_{it},x_{it}))^2 - \phi_{it}f(v_{it},x_{it})(\widehat{f}(v_{it},x_{it}) - f(v_{it},x_{it}))(\widehat{f}(x_{it}) - f(x_{it}))}{f^2(v_{it},x_{it})\widehat{f}(v_{it},x_{it})}.
\end{aligned}$$

From Silverman (1978) and Collomb and Hardle (1986), as $h \to 0$,

$$\sup |\widehat{f}(v_{it}, x_{it}) - f(v_{it}, x_{it})| = O_p[(N^{1-\varepsilon} h^{k+1})^{-\frac{1}{2}}],$$

$$\sup |\widehat{f}(x_{it}) - f(x_{it})| = O_p[(N^{1-\varepsilon} h^{k})^{-\frac{1}{2}}],$$

for any arbitrary small $\varepsilon > 0$.

Thus, under Assumption A.1, R_{it} is $O_p\left(\frac{1}{N^{1-\varepsilon}h^{k+1}}\right)$ and $\frac{1}{\sqrt{NT}}\sum_{t=1}^{T}\sum_{i=1}^{N} R_{it}$ is $O_p\left(\frac{1}{N^{1/2-\varepsilon}h^{k+1}}\right)$.

Under Assumption A.2, $\frac{1}{\sqrt{NT}}\sum_{t=1}^{T}\sum_{i=1}^{N} R_{it}$ is $o_p(1)$. So we could focus on the rest part.

From the Lemma above,

$$\frac{1}{\sqrt{NT}}\sum_{t=1}^{T}\sum_{i=1}^{N}\frac{\phi_{it}(\widehat{f}(x_{it}) - f(x_{it}))}{f(v_{it}, x_{it})} = \frac{1}{\sqrt{NT}}\sum_{t=1}^{T}\sum_{i=1}^{N}[E(\chi_{it}|x_{it}) - E(\chi_{it})] + o_p(1),$$

$$\frac{1}{\sqrt{NT}}\sum_{t=1}^{T}\sum_{i=1}^{N}\frac{\phi_{it} f(x_{it})(\widehat{f}(v_{it}, x_{it}) - f(v_{it}, x_{it}))}{f^2(v_{it}, x_{it})} = \frac{1}{\sqrt{NT}}\sum_{t=1}^{T}\sum_{i=1}^{N}[E(\chi_{it}|v_{it}, x_{it}) - E(\chi_{it})] + o_p(1).$$

Combined the results so far,

$$\begin{aligned}
& \frac{1}{\sqrt{NT}}\sum_{t=1}^{T}\sum_{i=1}^{N}\left(x_{it} - A_t'\pi_i\right)\left(\widehat{w}_{it} - E(w_{it}|x_{it}, \pi_i)\right) \\
= \; & \frac{1}{\sqrt{NT}}\sum_{t=1}^{T}\sum_{i=1}^{N}\left\{\frac{\phi_{it} f(x_{it})}{f(v_{it}, x_{it})} - \left(x_{it} - A_t'\pi_i\right)E(w_{it}|x_{it}, \pi_i) + [E(\chi_{it}|x_{it}) - E(\chi_{it})]\right. \\
& \left. -[E(\chi_{it}|v_{it}, x_{it}) - E(\chi_{it})]\right\} + o_p(1). \qquad (3.6.7) \\
= \; & \frac{1}{\sqrt{NT}}\sum_{t=1}^{T}\sum_{i=1}^{N}\left[\frac{\phi_{it} f(x_{it})}{f(v_{it}, x_{it})} - E(\chi_{it}|v_{it}, x_{it}) + E(\chi_{it}|x_{it}) - E(\chi_{it}|v_{it}, x_{it})\right] + o_p(1).
\end{aligned}$$

Since we assume that observations are i.i.d. across i, remember that

$$q_{it} = \frac{\phi_{it} f(x_{it})}{f(v_{it}, x_{it})} - E(\chi_{it}|v_{it}, x_{it}) + E(\chi_{it}|x_{it}) - E(\chi_{it}|v_{it}, x_{it}),$$

then

$$\begin{aligned} &\frac{1}{\sqrt{N}T} \sum_{t=1}^{T} \sum_{i=1}^{N} \left[\frac{\phi_{it} f(x_{it})}{f(v_{it}, x_{it})} - E(\chi_{it}|v_{it}, x_{it}) + E(\chi_{it}|x_{it}) - E(\chi_{it}|v_{it}, x_{it}) \right] \\ &\xrightarrow{d} N\left(0, var\left(\frac{1}{T}\sum_{t=1}^{T} q_{it}\right)\right). \end{aligned} \tag{3.6.8}$$

From equation (3.6.6), (3.6.7), and (3.6.8), we have

$$\sqrt{n}(\widehat{\beta} - \beta) \xrightarrow{d} N\left(0, \Delta var\left(\frac{1}{T}\sum_{t=1}^{T} q_{it}\right)\Delta'\right).$$

■

Bibliography

[1] Arellano, M., 2001, Panel data models: some recent developments, *Handbook of Econometrics*, Vol 5, Elsevier;

[2] Bai, J., 2003, Inferential theory for factor models of large dimensions, *Econometrica* 71, 135-171;

[3] Bai, J., 2009a. Panel data models with interactive fixed effects. *Econometrica* 77, 1229-1279;

[4] Bai, J., 2009b. Likelihood approach to small T dynamic panel models with interactive effects, working paper;

[5] Bai, J., and S. Ng, 2002, Determine the number of factors in approximate factor models, *Econometrica* 70, 191-221;

[6] Bai, J., and S. Ng, 2008, *Large Dimensional Factor Analysis*, Foundations and Trends in Econometrics, Vol 3, No 2, 89–163;

[7] Carro, J., 2007, Estimating dynamic panel data discrete choice models with fixed effects, *Journal of Econometrics* 140, 503–528;

[8] Chamberlain, G. 1980, Analysis of covariance with qualitative data, *Review of Economic Studies* 47, 225-238;

[9] Chamberlain, G. 2010, Binary response models for panel sata identification and information, *Econometrica* 78, 159–168;

[10] Collomb, G. and W. Hardle, 1986, Strong uniform convergence rates in robust nonparametric time series analysis and prediction: kernel regression estimation from dependent observations, *Stochastic Processes and Applications* 23, 77-89;

[11] Dong, Y., and A. Lewbel, 2012, Simple estimators for binary choice models with endogenous regressors, Working paper;

[12] Fernandez-Val, I., and M. Weidner, Individual and time effects in nonlinear panel data models with large N, T, working paper;

[13] Hayakawa, K., 2012, GMM estimation of short dynamic panel data models with interactive fixed effects, working paper;

[14] Honoré, B.E., and A. Lewbel, 2002, Semiparametric binary choice panel data models without strictly exogeneous regressors. *Econometrica* 70, 2053-2063;

[15] Hsiao, C., 2003, *Analysis of Panel Data*, Cambridge University Press;

[16] Hyslop, D., 1999, State dependence, serial correlation and heterogeneity in intertemporal labor force participation of married women, *Econometrica* 67, 1255 1294;

[17] Lewbel, A., 2000a. Semiparametric qualitative response model estimation with unknown heteroscedasticity or instrumental variables, *Journal of Econometrics* 97, 145-177;

[18] Lewbel, A., 2000b. Asymptotic trimming for bounded density plug-in estimators, unpublished manuscript;

[19] Lewbel, A., 2012. An overview of special regressor methods, working paper;

[20] Lewbel, A., Y. Dong, and T.T. Yang, 2012, Comparing features of convenient estimators for binary choice models with endogenous regressors, *Canadian Journal of Economics* 45, 809-829.

[21] Liang, Z., 2011, Binary response correlated random coefficient panel data models, working paper;

[22] Mundlak, Y. 1978, On the pooling of time series and cross section data, *Econometrica* 46, 69-85;

[23] Pesaran, M.H., 2006, Estimation and inference in large heterogeneous panels with a multifactor error structure, *Econometrica* 74, 967–1012;

[24] Newey, W. K. and D. McFadden (1994), Large sample estimation and hypothesis testing, in Handbook of Econometrics, vol. iv, ed. by R. F. Engle and D. L. McFadden, pp. 2111-2245, Amsterdam: Elsevier.

[25] Powell, J.L. J.H. Stock, and T.M. Stoker (1989), Semiparametric estimation of index coefficients, *Econometrica* 57, 1403-1430.

[26] Semykina, A., and J. Wooldridge, 2010, Estimating panel data models in the presence of endogeneity and selection, *Journal of Econometrics* 157, 375-380;

[27] Silverman, B. W. 1978, Weak and strong uniform consistency of the kernel estimate of a density and its derivatives, *Annal of Statistics* 6, 177-184;

[28] Wooldridge, J., 2010, *Econometric Analysis of Cross Section and Panel Data*, 2nd edition, MIT Press;

Acknowledgement

I am thankful to Arthur Lewbel, Stefan Hoderlein, and Zhijie Xiao for helping me build up my research skills. Their deep insights and thorough knowledge on econometric and economic issues frequently inspire me. I am also thankful to Jinghan Cai, Filippo De Marco, Bertan Turhan, Yatfung Wong for their helpful comments on my research. I thank Karim Chalak for his inputs.

Last but not least, I am greatful to my wife Ying Liao, my son Walter and my daughter Hannah for their endless supports.